Identifying Causal Directions from Text

Unsupervised Learning using Bayesian Framework

MACQUARIE
University
SYDNEY·AUSTRALIA

King Tao Jason Ng

Supervisor: Diego Mollá-Aliod

Associate Supervisors: Rolf Schwitter
Houying Zhu

School of Computing

A thesis submitted to Macquarie University for the degree of
Master of Research

October 2023

Dedication

To my wife, Joanna.

Declaration

Except where acknowledged in the customary manner, the material presented in this thesis is, to the best of my knowledge, original and has not been submitted in whole or part for a degree in any university.

King Tao Jason Ng
October 2023

Acknowledgements

I would like to thank Diego Mollá-Aliod and Rolf Schwitter for their advice and supervision, particularly in the areas of natural language processing and knowledge graphs. I would also like to thank Houying Zhu who offers feedback on Bayesian inference. I would like to thank members of Machine Learning Reading Group (MLRG) for insights into their research.

Abstract

Causality becomes increasingly important due to advances in artificial intelligence and machine learning, which demand robust interpretability and accountability. This significance is further underscored in the era of Large Language Models, like ChatGPT, where we witness achievements surpassing human performance in contextual understanding and the execution of text-related tasks. Even technology advancements, the importance of comprehending the data generation process that underlies causal directions becomes more evident, which can have valuable implications for various downstream tasks. In this project, we show empirically that word occurrences resemble the characteristics of causal directions. To achieve this, we determine a causal direction if its causal relation exists in the sentence. Identifying causal directions can add tremendous benefits into understanding semantics of the document. Firstly, knowing which entity is a cause and which one is an effect will help people understand complex phenomena more clearly. Secondly, being able to predict a likely outcome of certain events will help people make a better decision. Finally, many phenomena are difficult to comprehend in text. If Question Answering can extract causal relations and summarise them as causal directions, it will help people understand concepts easily. Nevertheless, there are two main challenges when identifying causal directions. Firstly, causal relations are few and far between in a document. Secondly, implicit causal relations make the task difficult. Hence, we propose a two-phase method: 1. Bayesian framework, which generates data from posteriors by incorporating word occurrences from the Internet's domains. 2. Bidirectional Encoder Representations from Transformers (BERT), which utilises semantics of words based on the context to perform classification. We have two scenarios. 1. Data augmentation, where word occurrences are integrated to expand the training data of the SemEval-2010 (Task 8) dataset. 2. Unsupervised learning, where no training data is provided. In this scenario, the proposed method learns from word occurrences. In data augmentation, the proposed method boosts an F1 score ever slightly compared with BERT, 94.34% versus 94.08%, but the difference is not statistically significant. BERT is used as the baseline. In unsupervised learning, where BERT or any other supervised methods cannot be used as a baseline, the proposed method performs significantly better than random guessing, achieving an F1 score of 49.10% versus 44.90%. Random guessing serves as the baseline for comparison. The study we carry out serves as a basis when we extend the proposed method to construct a network to capture multiple causal relations in future work.

Table of contents

List of figures **ix**

List of tables **xi**

1 Introduction **1**
1.1 Aims and Significance . . . 2
1.2 Research Questions . . . 2
1.3 Assumptions . . . 3
1.4 Thesis Outline . . . 4

2 Literature Review **5**
2.1 Causal Relation . . . 5
2.2 Datasets . . . 6
2.3 Causal Relation Detection . . . 8
2.3.1 Rule-Based . . . 8
2.3.2 Machine Learning . . . 9
2.3.3 Deep Learning . . . 10
2.3.4 Data Augmentation . . . 11
2.3.5 Unsupervised Learning . . . 11
2.4 Remarks . . . 12

3 Mathematical Background **13**
3.1 Probability and Probability Distribution . . . 13
3.1.1 Uniform Distribution . . . 15
3.1.2 Normal Distribution . . . 15
3.1.3 Other Distributions . . . 15
3.2 Directed Acyclic Graph . . . 15
3.3 Bayesian Inference . . . 16
3.3.1 Bayes' Rule . . . 17
3.3.2 Stan . . . 18
3.4 Further Reading . . . 19

4 Methodology **20**
4.1 Exploratory Analysis . . . 20
4.1.1 SemEval-2007 (Task 4) . . . 20
4.1.2 SemEval-2010 (Task 8) . . . 21
4.2 Bayesian Framework . . . 23
4.2.1 Hypothesis Testing . . . 23
4.2.2 Likelihoods . . . 24
4.2.3 Priors . . . 26
4.2.4 Posteriors . . . 29
4.2.5 Model Evaluation . . . 29
4.3 BERT . . . 31
4.4 Remarks . . . 32

5 Experiments **33**
5.1 Bayesian Framework . . . 34
5.1.1 Hypothesis Testing . . . 34
5.1.2 Likelihoods . . . 34
5.1.3 Priors . . . 35
5.1.4 Posteriors . . . 38
5.1.5 Model Evaluation . . . 38
5.2 BERT . . . 40
5.3 Results . . . 41
5.3.1 Summary . . . 46
5.3.2 Remarks . . . 48
5.4 Error Analysis . . . 49

6 Conclusion **52**
6.1 Discussion . . . 52
6.1.1 Limitations . . . 52
6.1.2 Likelihoods Redefinition . . . 53
6.2 Future Work . . . 54
6.2.1 C4 Dataset . . . 54
6.2.2 Earth Mover's Distance . . . 54
6.2.3 Mixture Models . . . 55
6.2.4 Bayesian Network . . . 55
6.2.5 Counterfactual . . . 56
6.3 Final Thoughts . . . 57

References **58**

Appendix A Methodology **63**
A.1 Exploratory Analysis 63

Appendix B Experiments **64**
B.1 Priors 64
B.1.1 $P(e1 \rightarrow e2)$ 65
B.1.2 $P(e2 \rightarrow e1)$ 70
B.2 Results 75
B.2.1 1(a) Data Augmentation — Bayesian Framework 75
B.2.2 1(b) Data Augmentation — BERT 87
B.2.3 1(c) Data Augmentation — Bayesian Framework + BERT 87
B.2.4 2(b) Unsupervised Learning — Bayesian Framework 89
B.2.5 2(c) Unsupervised Learning — Bayesian Framework + BERT 100

Appendix C Hardware and Software **102**
C.1 Code Repository 102
C.2 Hardware Requirements 102

Alphabetical Index **103**

List of figures

1.1 *The-chicken-or-the-egg* causal dilemma states that chickens hatch from eggs and eggs are laid by chickens. . . . 3

3.1 On average, there is a 20% chance of encountering a defective lightbulb. . . . 14
3.2 There is still a 20% chance of encountering a defective lightbulb, but it is less certain. 14
3.3 *e1* is a cause of *e2*. . . . 16
3.4 *e1* and *e2* show a causal relation, but no causal direction is specified. . . . 16
3.5 Since the spread is wider, it is less certain that $\mathbb{E}[e1 \rightarrow e2] = 0.7$ would hold. . . . 16
3.6 Because the spread is narrower, we are more certain that $\mathbb{E}[e1 \rightarrow e2] = 0.7$ would hold. 16
3.7 When the density centres around 0.9, the data supports $e1 \rightarrow e2$ to a large extent. . . 18
3.8 When the density centres mostly around 0, the data does not support $e1 \rightarrow e2$ much. . 18

4.1 Top 10 most mentioned entities in the training set. . . . 22

5.1 Both likelihoods $f(\mathbf{X} \mid \textit{death} \rightarrow \textit{suicide})$ and $f(\mathbf{X} \mid \textit{suicide} \rightarrow \textit{death})$ are shown respectively. . . . 35
5.2 When searching for the word *suicide*, 83,500 results are shown. . . . 35
5.3 Both priors $f(\textit{death} \rightarrow \textit{suicide})$ and $f(\textit{suicide} \rightarrow \textit{death})$ are equally likely. . . . 37
5.4 Both posteriors $f(\textit{death} \rightarrow \textit{suicide} \mid \mathbf{X})$ and $f(\textit{suicide} \rightarrow \textit{death} \mid \mathbf{X})$ are shown. . . . 38
5.5 Since the posteriors $f(\textit{accident} \rightarrow \textit{anxiety} \mid \mathbf{X})$ and $f(\textit{anxiety} \rightarrow \textit{accident} \mid \mathbf{X})$ heavily overlap, both *accident* $\rightarrow$ *anxiety* and *anxiety* $\rightarrow$ *accident* are likely. . . . 39
5.6 How the performance changes while varying the PPC ranges. . . . 42
5.7 The cost functions are presented based on whether PPC is enabled or disabled. The λ_r values are represented on the x-axis, while the y-axis corresponds to λ_e. In the top panel, the cost function $2\lambda_e + 293\lambda_r$ is displayed when PPC is enabled, while the bottom panel showcases the cost function $113\lambda_e + 43\lambda_r$ when PPC is disabled. . . . 43
5.8 How the performance changes while varying the PPC ranges. . . . 45

5.9 The displayed cost functions depend on whether PPC is enabled or disabled. The x-axis represents λ_e, while the y-axis represents λ_r. The top panel illustrates the cost function is $2\lambda_e + 321\lambda_r$ when PPC is enabled, while the bottom panel shows the cost function $125\lambda_e + 48\lambda_r$ when PPC is disabled. 45
5.10 Both $f(rain \rightarrow cancellation \mid \mathbf{X})$ and $f(cancellation \rightarrow rain \mid \mathbf{X})$ are shown. It is clear that the Bayesian model favours $cancellation \rightarrow rain$. 49
5.11 Both $f(moon \rightarrow perturbations \mid \mathbf{X})$ and $f(perturbations \rightarrow moon \mid \mathbf{X})$ are shown. It is clear that the Bayesian model favours $perturbations \rightarrow moon$. 50

6.1 If the light bulb market was flooded with light bulbs that had equally 20% as well as 30% chance of a defect, a probability distribution would look like multimodal. 55
6.2 $e1$, $e2$ and $e3$ show causal relations, but no directions are specified 56
6.3 $e3 \rightarrow e2$ is one possible way if $e3 \rightarrow e2$ exists. 56
6.4 $e3 \rightarrow e1 \rightarrow e2$ is another possible way if $e3 \rightarrow e2$ exists. 56

List of tables

2.1 Datasets are widely used for evaluating causal relation models. 8

4.1 Distributions of SemEval-2007 (Task 4) and SemEval-2010 (Task 8) are shown. . . . 20
4.2 The top 8 most mentioned entity pairs in the training set are displayed. 21
4.3 Sample data after pre-processing is shown. 31

5.1 All the Internet's domains used for the priors are shown. 36
5.2 Jeffreys' thresholds of evidence for BF are shown. 40
5.3 Thresholds that are used for the project are shown. 40
5.4 Cost function that aims to evaluate among Bayesian models is shown. 40
5.5 Results of experimental set-up 1(a) are displayed. 42
5.6 Results of experimental set-up 1(b) are displayed. 43
5.7 Results of experimental set-up 1(c) are displayed. 44
5.8 Results of experimental set-up 2(a) are displayed. 44
5.9 Results of experimental set-up 2(b) are displayed. 44
5.10 Results of experimental set-up 2(c) are displayed. 46
5.11 All the experimental set-ups are displayed. 48
5.12 Experimental values for sentence-source 8191 are shown. 50
5.13 Experimental values for sentence-source 8775 are shown. 51

B.1 How the performance changes when varying the PPC range (the 1^{st} run). 76
B.2 How the performance changes when varying the PPC range (the 2^{nd} run). 77
B.3 How the performance changes when varying the PPC range (the 3^{rd} run). 78
B.4 How the performance changes when varying the PPC range (the 4^{th} run). 79
B.5 How the performance changes when varying the PPC range (the 5^{th} run). 80
B.6 How the performance changes when varying the PPC range (the 6^{th} run). 81
B.7 How the performance changes when varying the PPC range (the 7^{th} run). 82
B.8 How the performance changes when varying the PPC range (the 8^{th} run). 83
B.9 How the performance changes when varying the PPC range (the 9^{th} run). 84
B.10 How the performance changes when varying the PPC range (the 10^{th} run). 85

B.11 Results of 1(a)(i) Data Augmentation — Bayesian (BF−, PPC−) are shown. 86
B.12 Results of 1(a)(ii) Data Augmentation — Bayesian (BF+, PPC+) are shown. 86
B.13 Results of 1(a)(ii) Data Augmentation — Bayesian (BF+, PPC−) are shown 87
B.14 Results of 1(b) Data Augmentation — BERT are shown. 87
B.15 Results of 1(c) Data Augmentation — Bayesian+BERT (BF+, PPC+) are shown. . . 88
B.16 Results of 1(c) Data Augmentation — Bayesian+BERT (BF+, PPC−) are shown. . . 88
B.17 How the performance changes when varying the PPC range (the 1^{st} run). 89
B.18 How the performance changes when varying the PPC range (the 2^{nd} run). 90
B.19 How the performance changes when varying the PPC range (the 3^{rd} run). 91
B.20 How the performance changes when varying the PPC range (the 4^{th} run). 92
B.21 How the performance changes when varying the PPC range (the 5^{th} run). 93
B.22 How the performance changes when varying the PPC range (the 6^{th} run). 94
B.23 How the performance changes when varying the PPC range (the 7^{th} run). 95
B.24 How the performance changes when varying the PPC range (the 8^{th} run). 96
B.25 How the performance changes when varying the PPC range (the 9^{th} run). 97
B.26 How the performance changes when varying the PPC range (the 10^{th} run). 98
B.27 Results of 2(b)(i) Unsupervised Learning — Bayesian (BF−, PPC−) are shown. . . . 99
B.28 Results of 2(b)(ii) Unsupervised Learning — Bayesian (BF+, PPC+) are shown. . . 99
B.29 Results of 2(b)(ii) Unsupervised Learning — Bayesian (BF+, PPC−) are shown. . . 100
B.30 Results of 2(c) Unsupervised Learning — Bayesian+BERT (BF+, PPC+) are shown. 100
B.31 Results of 2(c) Unsupervised Learning — Bayesian+BERT (BF+, PPC−) are shown. 101

Chapter 1

Introduction

> *"If I could sum up the message of this book in one pithy phrase, it would be that you are smarter than your data. Data do not understand causes and effects; humans do."*
>
> - Judea Pearl and Dana Mackenzie, *The Book of Why (Pearl and Mackenzie, 2018)*

One common form of questions in Question Answering (QA) is factoid questions such as *who is the* 30^{th} *prime minister of Australia?* in which answers can be found directly from a document. In contrast, answering causal questions are less explored because answers involve some kind of causal inference from sentences. In other words, answering causal questions involve identifying the relationship between events or entities where one causes another to take place. Singer et al. (1992) provide a great example: *Dorothy poured water on the fire. The fire went out*. Subsequently, if it is followed by a question *did she put out the fire?*, the answer will be *yes* because *poured water on* implies two sentences are causally linked. In the end, QA needs to analyse sentence semantics in order to find a cause-and-effect relation. The process in which such relation is established is called Causal Relation. In this project, we show empirically word occurrences resemble the characteristics of causal directions. Our aim is to gain insights into the data generation process that underlies causal directions, which can have valuable implications for various downstream tasks. To achieve this, we determine a causal direction if its causal relation exists in the sentence. Given two entities in the sentence that are known to have a causal relation, the causal direction tells us which one is a cause; which one is an effect. For instance, *poured water on* is a cause whereas *fire went out* is an effect. Therefore, the causal direction in this case is

$$\textit{poured water on} \longrightarrow \textit{fire went out}$$

There are two main challenges when identifying causal directions. Firstly, Gao et al. (2019) highlight causal relations are few and far between in a document, so insufficient data more likely makes predictions uncertain. Secondly, Zhao et al. (2021) argue implicit causal relations make the task difficult because not only a model is required to understand semantics, but also causal reasoning (Pearl and Mackenzie, 2018) in which machines are less capable of.

1.1 Aims and Significance

Ever since Pearl (2016) established a set of frameworks to tackle causal inference, causal inference has been a research topic across many disciplines such as economics, social science etc. However, the exploration of causal inference in Natural Language Processing (NLP) remains relatively limited (Girju, 2003; Oh et al., 2012; Verberne et al., 2007). Even if they do, most research has been focusing on identifying causal relations from sentences. Instead, what we do in this project is to determine a causal direction once its causal relation has been identified in the sentence. Identifying causal directions helps construct a Knowledge Graph (KG) which in turn facilitates reasoning (Kejriwal et al., 2021). Determining causal directions can also add tremendous benefits into understanding semantics of the document:

- Knowing which entity is a cause and which one is an effect will help us understand complex situations or phenomena more clearly. For instance, based on the article *Here's how scientists know the coronavirus came from bats and wasn't made in a lab* from the Conversation[1], are bats a cause to spread COVID-19 to humans?
- Being able to predict a likely outcome of certain events will help people make a better decision. For example, according to the article *Insulin and Diabetes* from Diabetes UK[2], should a GP prescribe insulin to help control diabetes?
- Many situations or phenomena are difficult to comprehend in text. If QA can extract causal directions and summarise them as a KG, it will help people understand a concept easily. For example, we can turn the following sentence, which is taken from Hendrickx et al. (2010):

  Suicide is one of the leading causes of death among pre-adolescents and teens, and victims of bullying are at an increased risk for committing suicide.

  into the corresponding KG:

  $$\textit{bullying} \longrightarrow \textit{suicide} \longrightarrow \textit{death}$$

  Thereafter, answering a *why* question such as *why do many pre-adolescents and teens commit suicide?* becomes easier as we simply perform graph inference on the KG.

In fact, these benefits are what have motivated the project.

1.2 Research Questions

The aim of this project is to answer the following research questions:

[1] https://theconversation.com/heres-how-scientists-know-the-coronavirus-came-from-bats-and-wasnt-made-in-a-lab-141850

[2] https://www.diabetes.org.uk/guide-to-diabetes/managing-your-diabetes/treating-your-diabetes/insulin

1. **Data Augmentation**: Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) has recently revolutionised the way we tackle NLP problems, including detecting causal directions. We want to answer this question — *Does the Bayesian framework that we use to augment data boosts performance for BERT?* The problem setting is as follows: Let us say two entities, $e1$ and $e2$, in the sentence that are known to have a causal relation. Our proposed method should return 1 if $e1$ causes $e2$; 0 if $e2$ causes $e1$. The model proposed for this research question will be compared when no data augmentation is applied.

2. **Unsupervised Learning**: The next question we want to answer is, *how well does the proposed method handle when no training is supplied?* Like the first question except this time no training data is provided. The Bayesian framework will solely learn from external sources. Most Machine Learning (ML) models perform poorly when the training data differs from the production data. This is so because when the production data comes from a different distribution, models may not generalise well. Unsupervised learning pushes to an extreme when no distributions can be learnt at all.

1.3 Assumptions

Causal relations can be found mostly either in a sentence or a document. On the other hand, multiple-document causal relations, where relations can span across different documents, are less often encountered. Determining multiple-document causal directions is arguably more difficult than a single document counterpart, for relevancy among documents must be first established. Multi-hop QA is recommended to manage this situation because QA can gather information from different parts of the documents to answer a question (Chen et al., 2019; Lu et al., 2022). However, multi-hop QA is out of scope for this project. Rather, we shall focus on a single causal relation occurring in the sentence. It is important to highlight bidirectional causal relations do occur sometimes, but they are not examined in this project. *The-chicken-or-the-egg* causal dilemma is a famous example as shown in Fig. 1.1.

$$egg \longleftrightarrow chicken$$

Fig. 1.1 *The-chicken-or-the-egg* causal dilemma states that chickens hatch from eggs and eggs are laid by chickens.

Additionally, we make the assumption that there are no latent, uncontrolled entities, or confounders that can affect both $e1$ and $e2$. In other words, we consider $e1$ and $e2$ as the sole entities available in the mention level. The objective of our studies is to explore how causal directions are reflected in the textual data at the mention level. Nevertheless, understanding the underlying causal relationship between two entities, such as *smoking* and *cancer*, which extends beyond mere mentions, would require controlled experimentation. It is important to note that our focus centers on Bayesian statistics, and we do not dive into the causality theory, as it lies beyond the scope of our research.

1.4 Thesis Outline

The thesis is divided into 6 chapters:

- Chapter 1 (i.e., this chapter) sets the scene by describing what causal relation and direction are. It also walks through challenges when identifying causal directions and potential benefits of the project. It mentions assumptions and research questions that we wish to answer.
- Chapter 2 provides a literature review, including how linguists define causal relations, how causal relations fit into NLP, datasets that people use to evaluate their models. It also examines some of the existing techniques for identifying causal relations.
- Chapter 3 provides a crash course on statistics. It first distinguishes two important concepts — probability and probability distribution, is followed by Directed Acyclic Graph (DAG), and finally explains what Bayesian inference is. You may want to skip this chapter and turn to next chapter should you already feel comfortable about statistics.
- Chapter 4 presents datasets used for our project. It also introduces the proposed method, particularly how we use the Bayesian framework to augment data. The concepts learnt from the previous chapter become useful to understand the proposed method.
- Chapter 5 details the nuts and bolts of the proposed method. It also shows how we set up experiments and results. Finally, it provides an error analysis.
- Chapter 6 is the final chapter where we discuss the results and explain some of the design choices. We also look at enhancements that can be made to improve performance.

Furthermore, Appendix A presents evidence that supports our methodology; Appendix B shows individual experimental results; Appendix C details hardware and software used in the project. Page 103 shows an index (or glossary).

Chapter 2

Literature Review

> *"What distinguishes language processing applications from other data processing systems is their use of knowledge of language."*

- Daniel Jurafsky and James H. Martin, *Speech and Language Processing (Jurafsky and Martin, 2008)*

Linguists study human languages. One area they do is sentence semantics in which relations between nouns and verbs in the sentences are closely examined. Unsurprisingly, one of the relations is causal relation.

2.1 Causal Relation

When a causal relation is expressed in sentences, a conjunction is usually encountered. English has numerous conjunctions to link phrases or clauses together. Some of them such as *because*, *as a result*, and *consequently* indicate an explicit causal relation because two parts of the sentence obviously express such relation. Indeed, Xuelan and Kennedy (1992) list down the conjunctions to express explicit causal relations, and some of the implicit causative verbs. Let us revisit the example from Chapter 1: *Dorothy poured water on the fire. The fire went out*. It is clear *poured water on* is a cause of *fire went out*. The causal relation in this case is implicit because the example does not use any conjunctions to link these events. Furthermore, Brinton and Brinton (2010) point out a word order sometimes represents the causal relation as shown below:

- *Dorothy became ill and left the party.*
- *Dorothy left the party and became ill.*

Both sentences have temporal relations. Nonetheless, in the first sentence we would interpret *she left the party because she became ill*. That is, *she became ill* is a cause of *she left the party*. The second one we suggest is merely a temporal relation. Having seen a couple of causal relation examples, we need to understand how linguists interpret causal relations. First, we introduce the concept of

Thematic Roles (Brinton and Brinton, 2010), which is the term used to describe the relationship between the different noun phrases and their respective roles in the sentence. For example, *Dorothy broke the vase.* According to thematic roles, *Dorothy* is the agent and *the vase* is the patient because the agent is an initiator whereas the patient is affected by the agent. Brinton and Brinton (2010) further explain some of the thematic roles:

- **Agent**: The entity that performs a certain action. It is a direct cause.
- **Patient**: The entity that is affected by an action. As a result, a state of the entity has changed.
- **Force**: Somewhat similar to the Agent, but it focuses on how willing or intentional the entity carries out an action.
- **Instrument**: A tool that is needed to carry out an action.

They insist that a causal relation involves either the force or agent role, and the patient role. For instance, *the sunlight yellowed the pages* in which *the sunlight* is the force role and *the pages* is the patient role. The same sentence can be re-written more clearly as *the sunlight caused the pages to become yellow*. A similar one is *Dorothy yellowed the pages* in which *Dorothy* is the agent and *the pages* is the patient role. Thus, they define a causal relation as changing the state of the entity. While this definition is satisfied from the linguistics' viewpoint, it leads to some situations where we would not normally consider causal relations. Let us illustrate what we mean with the previous example:

- *Dorothy poured water on the fire. The fire went out.*
- *Dorothy poured gasoline on the fire. The fire went out.*

The former is satisfied with general knowledge, but the latter is arguable because it contradicts common-sense knowledge as Singer et al. (1992) emphasize. The latter could still be considered as a causal relation provided we know how she used gasoline to put out the fire. Hence, only the former is viewed as the causal relation. In a nutshell, a causal relation, which changes the state of the entity in a sensible way, involves either the force or agent role, and the patient role.

2.2 Datasets

Before going any further, we summarise some of the datasets that are widely used for evaluating causal relation models.

- **SemEval-2007 (Task 4)**: The task is to categorise sentences into semantic relations such as *Cause-Effect*, *Instrument-Agency*, *Product-Producer* and so on (Girju et al., 2007). For the sake of the project, we only consider a particular category called Cause-Effect, which has 140 and 80 instances in the training and test sets respectively. 73 of 140 training instances are labelled as Cause-Effect in which 68 of them are $e2 \rightarrow e1$, where $e1$ and $e2$ are two entities marked in the instances. However, 41 of 80 test instances are considered as Cause-Effect in which all of 41 are

$e2 \rightarrow e1$. This dataset is respectable since SemEval is very established in semantic evaluation. Despite all the instances are roughly a sentence long, the dataset itself is rather small. Dunietz et al. (2015) notice causes and effects (i.e., $e1$ and $e2$) are restricted to noun clauses. They also comment some of the annotations rarely show causal relationships and suspect common-sense knowledge may not be noticeable to justify as Cause-Effect.

- **SemEval-2010 (Task 8)**: This dataset is an extension from the previous one (Hendrickx et al., 2010). Unlike the previous one, this is a multi-classification task as opposed to a binary classification in the SemEval-2007 (Task 4). In the interest of this project, we only consider a particular category called Cause-Effect, which has 1,003 and 328 Cause-Effect instances in the training and test sets respectively. 659 of 1,003 training instances are $e2 \rightarrow e1$; 194 of 328 test instances are $e2 \rightarrow e1$. Otherwise, both SemEval-2007 (Task 4) and SemEval-2010 (Task 8) share a lot of similarities. In fact, we argue the same comments from Dunietz et al. (2015) can be applied to this dataset (See Section 4.1.2 for more details).

- **Financial Document Causality Detection Shared Task (FinCausal 2020)**: This dataset focuses specifically on detecting causal relations in financial documents (Mariko et al., 2020). The shared task has two tasks: the first one is a binary classification where participants are asked to label whether the sentence is a causal relation. The second task is to extract chunks that are considered as causes and chunks that are considered as effects. In the interest of this project, we only dive into the second task, which has 1,109 and 641 instances in the training and test sets respectively. They also provide the evaluation set, which has 638 instances. Since causal or effect chunks vary, which in some cases they can be a sentence long, this task is noticeably harder than SemEval-2007 (Task 4) and SemEval-2010 (Task 8). Causal relations are defined based on their annotation scheme, which aligns with facts. Thus, this is more likely satisfied with causal relations compared with the previous two datasets.

- **Event Causality Identification with Causal News Corpus — Shared Task 3, CASE 2022**: This dataset also known as Causal News Corpus (CNC) is created for the shared task, which has two subtasks (Tan et al., 2022). The first is to label if a sentence contains causal relations; the second is to identify cause and effect spans given the sentence is labelled as causal relation. Since this shared task is like FinCausal 2020, we will not elaborate any further.

Table 2.1 summarises the above-mentioned datasets. Two additional datasets we have found are somewhat related to causal relations, but slightly off the project's course — SemEval-2012 (Task 7) and SemEval-2020 (Task 5). SemEval-2012 (Task 7) is open-domain common-sense reasoning (Gordon et al., 2012). Because of open-domain common-sense reasoning, it is expected additional sources of knowledge would be acquired for model building. SemEval-2020 (Task 5) is probably less relevant to our project, but arguably an important area in the causal relations (Yang et al., 2020). The task has two subtasks: the first one is to determine if a sentence is counterfactual; the second one is to identify antecedent and consequent phrases given the counterfactual sentence.

Dataset	Published Year	Number of Instances	Source	Availability
SemEval-2007 (Task 4)	2007	114	Wikipedia[a]	Public[b]
SemEval-2010 (Task 8)	2010	1,331	Wikipedia	Public[c]
FinCausal 2020	2020	2,388	Qwam[d]	License[e]
CNC Shared Task 3, CASE 2022	2022	3,559	News[f]	Public

Table 2.1 Datasets are widely used for evaluating causal relation models.

[a]https://www.wikipedia.org
[b]https://sites.google.com/site/semeval2007task4/data
[c]http://www.kozareva.com/downloads.html
[d]https://www.qwamci.com
[e]http://wp.lancs.ac.uk/cfie/fnp2020/
[f]https://github.com/tanfiona/CausalNewsCorpus

2.3 Causal Relation Detection

By now hopefully we have understood causal relations from the linguistics' perspective, it is a time to learn where causal relations fit into NLP. One discipline in NLP is referred as to Information Extraction (IE) where we distil information such as organisation names, dates, places and so on from the document. A prominent task that extracts proper names is called Named Entity Recognition (NER). Jurafsky and Martin (2008) provide an overview of NLP, particularly in areas of IE, NER and QA. Nevertheless, researchers are keen to obtain not only all the proper names, but also semantic relations among them. Unsurprisingly, one of them is causal relation.

Many causal relations are linked to entities such as the examples from Section 2.1. Therefore, causal relation identification can be found in either Relation Detection or Event Classification (Asghar, 2016). Many models have been proposed to identify causal relations. In fact, Reimann (2021); Yang et al. (2021) categorise the models based on the techniques into three categories: *Rule-Based*, *Machine Learning* and *Deep Learning*. And we have extended these to cover *Data Augmentation* and *Unsupervised Learning*. It is possible a model can fall into multiple categories. For example, if a model follows a rule-based approach without using any training set, it can be categorised under both the Rule-Based and Unsupervised Learning.

2.3.1 Rule-Based

Rule-based approaches rely on linguistic or syntactic patterns. These often need to study linguistic extensively in order to identify causal relations. For instance, KHOO et al. (1998) look for explicit causal relations in the articles from Wall Street Journal using causal words such as *because*, *that's why*, *the result was* and so on, as well as search for the relations using syntactic patterns like *Verb-Noun-Phase-Adjective*, *if-then* and others. Their model achieves a 68% recall, but a substantial number of

errors is made due to complex sentence structures. One shortcoming is that it is not able to manage implicit causal relations.

2.3.2 Machine Learning

To mitigate implicit discourse relations, Marcu and Echihabi (2002) build an extensive training data by leveraging discourse markers such as *because* or *but*. They then utilize the data to train a Naive Bayes classifier. Similarly, Hidey and McKeown (2016) develop a model by leveraging parallel Wikipedia articles and creating a collection of alternative lexicalizations called AltLex[1]. Subsequently, they employ Support Vector Machine (SVM) to detect implicit causal relations. Their methodology achieves a 75.33% F1 score and 85.85% accuracy. While their model yields promising results, creation of the required corpus involves substantial efforts. To eliminate the need for corpus creation, co-occurrences is proposed to measure the strength of two causal entities (Kroeger, 2005). If a pair of entities forms the causal relation, it is anticipated these entities will often be mentioned together. One notable method to measure co-occurrence is Pointwise Mutual Information (PMI) (Glickman et al., 2005). Let us say two entities, $e1$ and $e2$. Suppes (1973) points out $e1$ is a possible cause of $e2$ if $e2$ is mentioned more frequently with $e1$ than on its own.

$$P(e2 \mid e1) > P(e2) \tag{2.1}$$

We rewrite Equation (2.1) as follows:

$$\frac{P(e2 \cap e1)}{P(e1)P(e2)} > 1 \tag{2.2}$$

Equation (2.2) is elegant if $e1$ and $e2$ establish the causal relation, but it fails to determine its causal direction. For example, if $e2$ is a cause of $e1$, we have

$$P(e1 \mid e2) > P(e1) \tag{2.3}$$

After a couple of algebraic manipulations, we end up

$$\frac{P(e1 \cap e2)}{P(e2)P(e1)} > 1 \tag{2.4}$$

Equation (2.2) and (2.4) are now identical. That is, we cannot distinguish $e1 \rightarrow e2$ from $e2 \rightarrow e1$ using PMI. Nevertheless, PMI is often used in causal relation identification. For instance, Moghimifar et al. (2020) use PMI to measure the strength of any pairs of words before constructing causal Bayesian networks (See Section 6.2.4). In a similar vein, Do et al. (2011) calculate distributional similarity of event co-occurrences based on PMI to identify causal relations. Using the co-occurrence of lexical pairs, Chang and Choi (2005) propose a probabilistic method to extract causal relations by

[1] AltLex refers to alternative words or phrases that convey a similar or sometimes the same concept with different linguistic words.

employing cue phrase and lexical pair probabilities. Since they do not access to the causal relation annotated corpus, they use the Expectation-Maximization (EM) algorithm to estimate parameters of prior probability, cue phrase probability, and lexical pair probability. The EM algorithm estimates a latent variable through an iterative optimization algorithm (Murphy, 2013). While the EM algorithm can incorporate prior knowledge, it does so by setting initial parameter values only. If we have no prior knowledge about the latent variable, the EM algorithm is recommended due to computational efficiency. On the contrary, if we do have prior knowledge, Bayesian inference is often the best approach because it provides a mathematical framework to integrate prior knowledge by computing prior distributions (Lambert, 2018).

2.3.3 Deep Learning

Before the emergence of pre-trained language models like BERT, ChatGPT[2], and GPT-4, Recurrent Neural Networks (RNNs) were widely used for identifying causal relations. For instance, Dasgupta et al. (2018) propose a bidirectional Long Short-Term Memory (LSTM), which is a type of RNNs. They combine feature vectors to extract causal relations in which they pull causes, effects, and causal connectives from various datasets. While their model serves as a benchmark, it is challenging to compare it due to a different evaluation method.

With the focus shifting to BERT, we present our thoughts in this area. BERT, ChatGPT, and GPT-4 utilize a pre-training technique where models are trained on extensive data and then fine-tuned for specific applications. BERT's dominance in the field of language models can be attributed to its context-based embedding model, which enables it to understand word meanings based on contextual cues (Devlin et al., 2019). For instance, to find causal relations, Khetan et al. (2021) test out three architectures, which are C-BERT, Event Aware C-BERT and Masked Event C-BERT, on top of the pre-trained BERT. They find Event Aware C-BERT has slight performance gain compared with the other two. Moreover, Li et al. (2021) incorporate causal knowledge into BERT and fine-tune BERT using minimal supervision. Additionally, the Graph Convolutional Network (GCN) (Tian et al., 2021; Tran Phu and Nguyen, 2021) has also gained popularity due to its graphical resemblance to a Directed Acyclic Graph (DAG) (See Section 3.2), making it easier for humans to interpret causal relations. The existing models except Dasgupta et al. (2018) introduced thus far focus on the causal relation identification as opposed to the determination of causal directions. In contrast, Hosseini et al. (2021) address the task by predicting the directions of causal pairs in textual content. They examine two bidirectional transformer-based language models, namely BERT and SpanBERT, which they find SpanBERT achieves superior performance compared with BERT.

[2]GPT stands for Generative Pre-trained Transformer.

2.3.4 Data Augmentation

Human beings unsurprisingly make inferences from just a small amount of data. Griffiths and Tenenbaum (2009) call such process as Causal Induction. To understand causal induction, let us demonstrate it using two hypothetical events, $e1$ and $e2$:

- $e1$: *A man speeded at 180km per hour in the Hume Highway.*
- $e2$: *Five people including a child injured in a head-on collision in the Hume Highway.*

Without a broader context, a sensible causal relation in this hypothetical example is $e1 \rightarrow e2$. However, it is possible but arguably less likely the opposite would hold true. That is, the man suddenly speeded after he saw five people injured in the Hume Highway. Many language models would be difficult to reason due to not enough instances being provided. In fact, one thing all the above-mentioned datasets share is a limited sample size due to expensive annotations. To address this limitation, data augmentation is proposed. Many language models handle zero-shot or few-shot learning by augmenting data[3]. For instance, Li et al. (2021), whom we discussed in Section 2.3.3, extract causal knowledge from CausalBank, ConceptNet etc. Nevertheless, language models are in general less successful in zero-shot compared with few-shot learning. Wei et al. (2021) show how to fine-tune the language models via instructions to improve the performance for the zero-shot scenario.

2.3.5 Unsupervised Learning

ML methods consists of three categories: *Supervised*, *Unsupervised*, and *Reinforcement Learning*. All the methods introduced in Section 2.3 fall under supervised learning, with the exception of KHOO et al. (1998), which follows a rule-based approach. Supervised learning has a key advantage — Reliability. As noted by Gharagozlou et al. (2023), it can effectively learn from ground truth data. However, this approach has several limitations, including the cost of annotations, the challenge of creating sufficiently large datasets, and the time required for data preprocessing. In contrast, unsupervised learning can address some of these challenges by leveraging knowledge bases or external data sources. To reduce the demand for annotation, Ittoo and Bouma (2013) introduce a minimally-supervised model that relies on a few seed examples of causal relations. This model leverages an open-domain corpus to extract additional causal relations. While this approach mitigates the shortage of annotated data, it is not flexible enough to handle various types of causal relations. Unsupervised methods are more flexible to discover patterns that might not have been present in the seed examples. As introduced in Section 2.3.2, Do et al. (2011) take a different approach. They compute event co-occurrences from an unannotated corpus, leveraging distributional similarity as a proxy to identify causal relations. Although their paper is titled *Minimally Supervised Event Causality Identification*, we would classify it as unsupervised learning rather than minimally supervised. This is because

[3]Few-shot learning is that a model is trained on a handful of labelled data; zero-shot learning is that a model is trained on some classes to recognise other classes of data. However, in the unsupervised learning, models are not provided any labelled data.

computing event co-occurrences does not require annotated data. However, they do annotate articles collected from Cable News Network (CNN) to evaluate their model.

2.4 Remarks

While deep learning is capable of learning data representations, it relies on having a substantial amount of data, which can be challenging for causal relations as they are not frequently encountered in documents. On the other hand, ML methods can mitigate the need for annotated data but may not achieve the same performance as deep learning. Ideally, a method that combines the strengths of both approaches would be desirable. Many data augmentation and unsupervised learning approaches leverage knowledge bases. Therefore, the choice of knowledge base is crucial as an inappropriate selection can lead to substandard performance. Furthermore, since predictions always involve some level of uncertainty, it is necessary to have a mathematical framework to quantify this uncertainty.

Chapter 3

Mathematical Background

"Uncertainty arises because of limitations in our ability to observe the world, limitations in our ability to model it, and even because of innate nondeterminism."

- Daphne Koller and Nir Friedman, *Probabilistic Graphical Models: Principles and Techniques (Koller and Friedman, 2009)*

Before going any further, it is essential to introduce mathematical notations. We use an upper case such as D or E as a random variable, which assigns a numerical value to each possible outcome of an experiment, or statistically called a sample space. The exception is C, which is reserved for a count function. For instance, $C(apple)$ is how many times the word *apple* appears in a document. A lower case such as c or d is a value that the random variable can take except e, which is denoted as an event or entity. For example, if we have two entities, we write them as $e1$ and $e2$. We use boldface like $\mathbf{X}$ as a matrix. Greek letters such as θ or μ have their own meanings in statistics. However, we introduce them as we go along. We reserve P as a probability and f as a probability distribution. This chapter reviews materials in statistics. You may want to skip this chapter and turn to Chapter 4 should you already know fundamental concepts in statistics. Speaking of probability and probability distribution, we now explain these two somewhat related, but distinct concepts in the next section.

3.1 Probability and Probability Distribution

Probability is a numerical value that indicates the likelihood of an event occurring. It is represented as a value between 0 and 1, where 0 is impossible to occur and 1 is a definite event. For example, we can define a random variable D that takes on the value 1 if a light bulb is defective or 0 if it is not, denoted as corresponding probability $P(D = 1)$ and $P(D = 0)$ respectively. Let us say a probability of finding a defect in a light bulb from a specific manufacturer is $P(D = 1) = 0.2$, which means one out of every five light bulbs is defective. On the other hand, a probability distribution is a function that

describes the likelihood of different possible values that a random variable can take[1]. For instance, the probability distribution of finding a defect in a light bulb across all manufacturers can be represented by a function $f(P(D))$, where $P(D)$ is a random variable that takes on values like 0.1, 0.2, 0.3, and so on[2]. If we assume that the likelihood of defects follows a normal distribution with the mean of 0.2, the probability distribution might look like the one shown in Fig. 3.1. In this case, the probability $P(D) = 0.2$ would be more likely to occur than others[3]. In simple terms, if you randomly pick up a light bulb repeatedly, on average, there is a 20% chance of encountering a defective one. Another interpretation is that light bulbs with a 20% defect rate dominate the light bulb market.

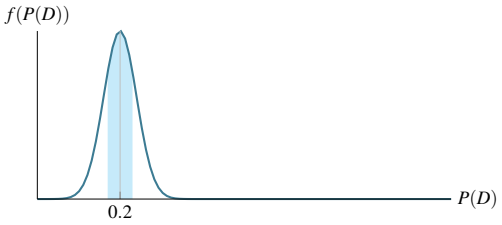

Fig. 3.1 On average, there is a 20% chance of encountering a defective lightbulb.

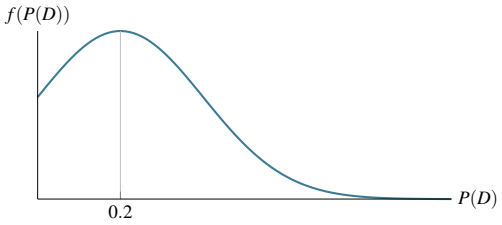

Fig. 3.2 There is still a 20% chance of encountering a defective lightbulb, but it is less certain.

If we want to determine the probability density[4] between any two values, such as $0.15 \leq P(D) \leq 0.25$, it corresponds to the light blue area in Fig. 3.1. The larger the light blue area, the more likely the value of $P(D)$ falls within that range. Now, let us modify the standard deviation of the same probability distribution, as illustrated in Fig. 3.2. When randomly picking up a light bulb again, there is still a 20% chance of encountering a defective one. However, due to the wider spread of the distribution, it is also likely to select a light bulb with a 10% or 30% chance of defect. In simple terms, the variance of the probability distribution governs the level of uncertainty in the selection process. Both the mean, which is denoted as $\mathbb{E}$, and variance can be computed for most probability distributions. Before we conclude this section, it is crucial to emphasize a fundamental characteristic of probability distributions. When we have a probability distribution and its corresponding parameter(s), we have the ability to simulate samples. For instance, a normal distribution requires two parameters, μ and σ, which will be explained shortly in Section 3.1.2. By providing specific values for these parameters, we can generate samples from the distribution.

In summary, a probability represents a single value indicating the likelihood of an event occurring, whereas a probability distribution is a function that describes the likelihood of all possible outcomes

[1] In the frequentist perspective, probability is based on observed data and is interpreted as the long-run relative frequency of events in repeated trails. However, the Bayesian interprets probability as a measure of our degree of belief or uncertainty and allows for domain knowledge encoded in priors. Hence, Bayesian can define probability distributions for all possible events including subsets (i.e., those that have not been directly observed) (Lambert, 2018).

[2] Although we present them as discrete values for illustrative purposes, probabilities in this case are actually continuous.

[3] We have simplified $P(D = 1)$ as $P(D)$ from now on.

[4] In statistics, density refers to the probability distribution of a random variable.

for a random variable. Thus far, we have not yet defined the types of probability distributions, so it is the time to introduce some of well-known probability distributions.

3.1.1 Uniform Distribution

If a sample space consists of a finite number of outcomes and each outcome is equally likely, we can represent the probability distribution using a uniform distribution, which is often considered as an uninformative distribution. When we say that $P(D)$ follows a uniform distribution, it can be represented as:

$$P(D) \sim Uniform(a,b). \tag{3.1}$$

Here, a and b are parameters that define the range of possible outcomes. In our case, since $P(D)$ is bound between 0 and 1, $a = 0$ and $b = 1$. The mean of the uniform distribution can be calculated as:

$$\mathbb{E} = \frac{a+b}{2} = 0.5. \tag{3.2}$$

Put it simply, if the probability distribution follows a uniform distribution, it means that all outcomes within the specified range have an equal chance of occurring. The mean of the distribution is calculated by taking the average of the lower and upper bounds of the range.

3.1.2 Normal Distribution

The Normal or Gaussian distribution is very important in statistics. If $P(D)$ follows a normal distribution, it is

$$P(D) \sim N(\mu,\sigma), \tag{3.3}$$

where μ is the mean and σ is the standard deviation.

3.1.3 Other Distributions

Student's t-Distribution, Cauchy distribution, Exponential distribution, Gamma distribution, Inverse-gamma distribution and Log-normal distribution are also used in this project. Due to the space constraint, more details can be referred to classical statistics textbooks (See Section 3.4).

3.2 Directed Acyclic Graph

Let us re-visit the example from Section 2.1: *Dorothy poured water on* ^{e1} *the fire. The fire went out* ^{e2}. The same relation can be presented graphically in Fig. 3.3.

Fig. 3.3 *e1* is a cause of *e2*.

Indeed, Pearl (2016) calls this diagram a Directed Acyclic Graph (DAG) where nodes are events and edges are causes[5]. Because a graphical model like DAG is intuitive to follow, it has recently become a predominant toolbox to tackle event causality identification (Tran Phu and Nguyen, 2021). To illustrate conceptually how a DAG is constructed to represent the causal direction like Fig. 3.3, let the diagram shown in Fig. 3.4 be an underlying structure. If the probability of $e1 \rightarrow e2$ is higher than $e2 \rightarrow e1$, we will conclude $e1 \rightarrow e2$ is the correct structure. If both $e1 \rightarrow e2$ and $e2 \rightarrow e1$ had roughly the same probability, we would attempt to conclude both $e1 \rightarrow e2$ and $e2 \rightarrow e1$ hold. As a DAG is a directed graph without any cycles, this situation will be statistically invalid despite this can happen.

Fig. 3.4 *e1* and *e2* show a causal relation, but no causal direction is specified.

So far, we have somehow known all these probabilities, but in practice how do we compute them? To build upon the assumption of no directed cycles, Bayesian inference hopefully sheds some light on this problem.

3.3 Bayesian Inference

Statistical inference is a framework we use to draw an inference from a sample of data about some aspects of a population. Two mainstreams of philosophy behind statistical inference are named: Frequentist and Bayesian. The Frequentist statistics treat parameters, we shall now call θ, as fixed but unknown[6]. Let us revisit Fig. 3.3 and θ be $e1 \rightarrow e2$. We assume $\mathbb{E}[e1 \rightarrow e2] = 0.7$ would quantify $e1 \rightarrow e2$. That is, on average 70% chance the statement $e1 \rightarrow e2$ holds. Because no information is provided respecting uncertainty, both scenarios shown in Fig. 3.5 and Fig. 3.6 would equally satisfy $\mathbb{E}[e1 \rightarrow e2] = 0.7$.

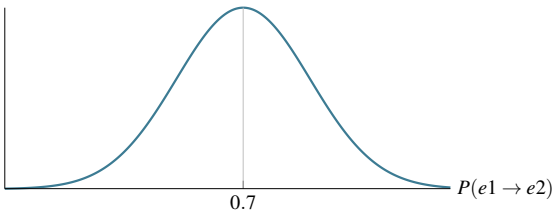

Fig. 3.5 Since the spread is wider, it is less certain that $\mathbb{E}[e1 \rightarrow e2] = 0.7$ would hold.

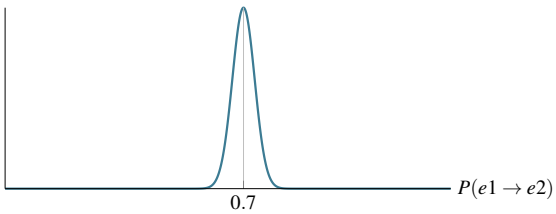

Fig. 3.6 Because the spread is narrower, we are more certain that $\mathbb{E}[e1 \rightarrow e2] = 0.7$ would hold.

[5]When we say $e1 \rightarrow e2$, which is taken from the Fig. 3.3, we mean it is likely e1 causes e2 because of uncertainty involved.

[6]In statistics, θ represents a parameter of interest that we want to find out from data.

If we had to choose one based on certainty, we would normally incline toward Fig. 3.6 since the spread is narrower. This represents high certainty around the mean of the normal distribution. Moreover, we have no way to know if 0.7 is indeed a correct probability[7] to qualify $e1 \rightarrow e2$ because this number is almost always computed from a sample, not a population. That is, this probability is likely different for another sample. However, Bayesian takes a radical approach in which we assume θ can vary. We argue such assumption is reasonable. Unless we know the data generating process, every value of θ is possible although some are more likely. The fact θ can vary will give us a probability distribution (Lambert, 2018).

3.3.1 Bayes' Rule

Bayesian inference uses Bayes' rule as in Equation (3.4) to compute a probability distribution of θ, which is denoted as a random variable of interest, given the data $\mathbf{X}$. We need to point out that every term except $f(\mathbf{X} \mid \theta)$ and $f(\mathbf{X})$ in the equation is now a probability distribution as opposed to a probability[8] (See Section 3.1). Before we collect the data, we have a probability distribution of our belief, or statistically speaking called a prior. If a certain value of θ holds, a likelihood weighs a probability of the data to be seen. The Bayes' rule simply provides a formula of how a prior and a likelihood ought to be put together to derive a posterior:

$$\overbrace{f(\theta \mid \mathbf{X})}^{Posterior} = \frac{\overbrace{f(\mathbf{X} \mid \theta)}^{Likelihood}\overbrace{f(\theta)}^{Prior}}{\underbrace{f(\mathbf{X})}_{Denominator}} \tag{3.4}$$

In practice, we often omit the denominator, $f(\mathbf{X})$, because computing it would require tremendous efforts and it is also a constant when ranking. Therefore, the simplified version becomes:

$$f(\theta \mid \mathbf{X}) \propto f(\mathbf{X} \mid \theta) f(\theta) \tag{3.5}$$

[7] Statisticians often make an assumption to simplify analysis, but sometimes we need data to tell us if the assumption is valid.

[8] A likelihood, which is rather called a likelihood function, is not a probability distribution because the density does not sum to 1 (Lambert, 2018).

Likelihoods

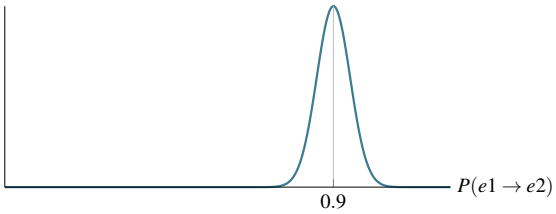

Fig. 3.7 When the density centres around 0.9, the data supports $e1 \rightarrow e2$ to a large extent.

Fig. 3.8 When the density centres mostly around 0, the data does not support $e1 \rightarrow e2$ much.

The task is to find out how likely the statement $e1$ *is a cause of* $e2$ holds. Let θ be $e1 \rightarrow e2$. If $e1 \rightarrow e2$ holds in the data, the likelihood $f(\mathbf{X} \mid e1 \rightarrow e2)$ tells us how likely the data will be encountered. For example, if the density of $f(\mathbf{X} \mid e1 \rightarrow e2)$ centres mostly around 0.9, as shown in Fig. 3.7, the data supports the statement to a large extent. For the sake of simplicity, we could think as if $\mathbb{E}[f(\mathbf{X} \mid e1 \rightarrow e2)] = 0.9$. Nonetheless, whilst the density of $f(\mathbf{X} \mid e1 \rightarrow e2)$ centres around 0 as displayed in Fig. 3.8, or $\mathbb{E}[f(\mathbf{X} \mid e1 \rightarrow e2)] = 0$, the data does not seem to support the statement much.

Priors

When we have sufficient data, likelihoods will probably be good enough to tell a whole story. Often, a lack of data would prompt us to look for an alternative way to model the problem. One way to do so is priors. Priors allow us to state our beliefs before collecting the data, but beliefs must be represented in probability distributions. A belief is often subjective. In fact, what one believes could be completely disagreed by others. Hence, it is important to state priors explicitly. In our case, $f(e1 \rightarrow e2)$ is the prior.

Posteriors

The aim of Bayesian inference is to compute the posterior, $f(e1 \rightarrow e2 \mid \mathbf{X})$, via the Bayes' rule as in Equation (3.4). The posterior distribution summarises the uncertainty over $e1 \rightarrow e2$.

3.3.2 Stan

When prior and posterior distributions are from the same probability family[9], we say the prior is a conjugate prior for the likelihood function. When this happens, we can work out the posterior analytically. In other situations, we use the statistical programming language called Stan to approximate the posterior distribution (Kruschke, 2015; Lambert, 2018). Stan is a powerful tool commonly used for Bayesian data analysis, and it enables us to perform complex computations and obtain posteriors in situations where analytical solutions are not readily available.

[9] A probability family is a set of probability distributions that have a common mathematical structure (Kruschke, 2015).

3.4 Further Reading

Since this chapter serves as a crash course in statistics, it is impossible to cover every aspect of DAG and Bayesian inference. However, we provide a couple of textbooks that are worthwhile to read:

- *Mathematical Statistics with Applications* (Wackerly et al., 2002): This textbook provides an introduction to statistics, including some of the well-known distributions.
- *Causal Inference in Statistics a Primer* (Pearl, 2016) and *The Book of Why* (Pearl and Mackenzie, 2018): Judea Pearl discusses DAG from the causal inference perspective.
- *A Student's Guide to Bayesian Statistics* (Lambert, 2018), *Doing Bayesian Data Analysis* (Kruschke, 2015) and *Statistical Rethinking* (McElreath, 2015): These three textbooks are considered as an introductory to Bayesian statistics. In fact, our methodology, as we will walk through in the next chapter, is based on these three books.

Chapter 4

Methodology

"Sometimes classical statistics gives up. Bayes never gives up... so we're under more responsibility to check our models."

- Andrew Gelman, *http://www.stat.columbia.edu/∼gelman/book/gelman_quotes.pdf*

This chapter introduces the dataset used in the project and follows by the proposed method.

4.1 Exploratory Analysis

Having briefly discussed the datasets in Section 2.2, we decide to use two of them — SemEval-2007 (Task 4) and SemEval-2010 (Task 8) — to evaluate the proposed method. Because our purpose is to determine causal directions, the two datasets are filtered to Cause-Effect only. Table 4.1 provides an overview of both datasets considered in the project, including the distributions between training and test splits.

Datasets		**SemEval-2007 (Task 4)**		**SemEval-2010 (Task 8)**	
		Raw Count	Percentage	Raw Count	Percentage
Training	$e1 \rightarrow e2$	5	6.85%	344	34.30%
	$e2 \rightarrow e1$	68	93.15%	659	65.70%
	Total	73	100.00%	1,003	100.00%
Test	$e1 \rightarrow e2$	0	0.00%	134	40.85%
	$e2 \rightarrow e1$	41	100.00%	194	59.15%
	Total	41	100.00%	328	100.00%

Table 4.1 Distributions of SemEval-2007 (Task 4) and SemEval-2010 (Task 8) are shown.

4.1.1 SemEval-2007 (Task 4)

The Cause-Effect data is divided into training and test sets. The training one has 73 Cause-Effect instances, but only 5 of them are $e1 \rightarrow e2$. Nevertheless, all 41 Cause-Effect instances in the test

set are $e2 \rightarrow e1$. Each instance starts with an identifier and is followed by a sentence in which two entities are marked by <e1> and <e2>. If a sentence is labelled as Cause-Effect, the two entities also tell us the causal direction. Example 4.1 shows $e2$ is a cause of $e1$[1].

```
1 sentence-source = 9, Sentence = "People in Hawaii might be feeling <e1>
    aftershocks</e1> from that powerful <e2>earthquake</e2> for weeks."
2 WordNet(e1) = "aftershock", WordNet(e2) = "earthquake", Cause-Effect(e2,e1) =
     "true", Query = "aftershocks from *"
```

Example 4.1 sentence-source: 9

The reason is, aftershocks are earthquakes, but on a smaller scale after a major one. After carefully examining this data, we have decided to drop this dataset due to directions in the test set are all one-sided.

4.1.2 SemEval-2010 (Task 8)

SemEval-2010 (Task 8) has 1,003 instances labelled as Cause-Effect in the training set. 659 of those are $e2 \rightarrow e1$. In the test data, 328 instances are Cause-Effect, but 134 of them are $e1 \rightarrow e2$. A sentence is considered as Cause-Effect if two entities, which are labelled as <e1> and <e2>, show a causal relation. Example 4.2 shows $e1$ is a cause of $e2$ (This example will in fact be used for illustrative purposes throughout Chapter 5).

```
1 sentence-source = 27  "<e1>Suicide</e1> is one of the leading causes of <e2>
    death</e2> among pre-adolescents and teens, and victims of bullying are
    at an increased risk for committing suicide."
2 Cause-Effect(e1,e2) Comment:
```

Example 4.2 sentence-source: 27

Frequency	Entities Pair
4	*collision* $\rightarrow$ *fire*
3	*death* $\rightarrow$ *grief*
3	*earthquake* $\rightarrow$ *fire*
3	*storm* $\rightarrow$ *damage*
3	*washing* $\rightarrow$ *shrinkage*
3	*bacteria* $\rightarrow$ *infection*
3	*spilling* $\rightarrow$ *burn*
3	*injury* $\rightarrow$ *discomfort*

Table 4.2 The top 8 most mentioned entity pairs in the training set are displayed.

A critical step we perform is to validate there are no bidirectional causal relations in the dataset. Thankfully, they do not exist. We list down the top 8 mostly mentioned entity pairs in Table 4.2.

[1] sentence-source is an identifier.

As for descriptive statistics, Fig. 4.1 shows the top 10 mostly mentioned entities in the training set. Nevertheless, we do have some comments about the training set.

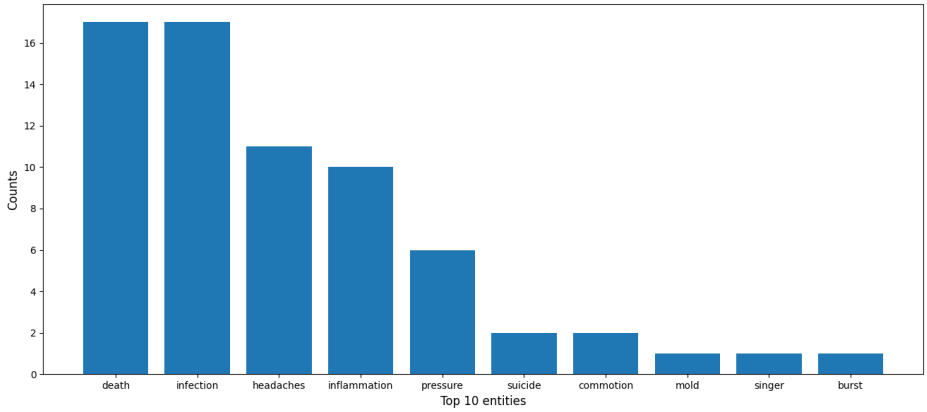

Fig. 4.1 Top 10 most mentioned entities in the training set.

Comments

As we alluded in Section 2.2, these are comments we want to highlight. Firstly, annotations may be right, but some sentences are not compatible with common sense knowledge (See Example 4.3).

```
sentence-source = 71  "The continuing Nigerian <e1>outbreak</e1> is the
    biggest ever caused by the <e2>vaccine</e2>."
Cause-Effect(e2,e1) Comment:
```

Example 4.3 sentence-source: 71

This says *vaccine* is a cause of *outbreak*, but vaccines are designed to prevent outbreaks. It is virus that causes an outbreak. Secondly, some do not have any causal relations (See Example 4.4).

```
sentence-source = 296 "This <e1>meeting</e1> establishes the <e2>laws</e2>
    which govern the church and the priorities for the coming year."
Cause-Effect(e1,e2)
Comment: "laws" is a metonymy for "rule of law", which is a state
```

Example 4.4 sentence-source: 296

This states *meeting* is a cause of *laws*, but we suspect it is more a social issue that causes the laws to be established. Finally, a mediator (Pearl and Mackenzie, 2018) is missing in some instances. When one entity causes another one via a third entity, the third entity is called a mediator (See Example 4.5).

```
sentence-source = 963 "A first <e1>revolution</e1> was triggered by the
    growing use of reading and <e2>writing</e2>."
Cause-Effect(e2,e1) Comment:
```

Example 4.5 sentence-source: 963

This implies increasing literacy through reading and writing would allow people to acquire knowledge which in turn led to a revolution. Hence, we argue writing is a factor, but not necessarily the direct cause to the revolution. Finally, we have also surveyed on the test set. 20 samples from the test set are taken and we find 2 of them (i.e., 10%) are dubious (See Section A.1 for more details).

4.2 Bayesian Framework

The proposed method consists of two phases — Bayesian framework and BERT. Let us start with the Bayesian framework. If an instance is labelled as Cause-Effect, the task is to determine a causal direction. Provided two entities, namely $e1$ and $e2$, the direction will be either $e1 \rightarrow e2$ or $e2 \rightarrow e1$. First, we set up two models as follows:

$$\text{Model 1: } f(e1 \rightarrow e2 \mid \mathbf{X})$$
$$\text{Model 2: } f(e2 \rightarrow e1 \mid \mathbf{X})$$

where $\mathbf{X}$ is the training set and f is a probability distribution. Given the training set, we compute the probability distribution of $e1 \rightarrow e2$ for Model 1, and similarly $e2 \rightarrow e1$ for Model 2.

4.2.1 Hypothesis Testing

To formulate the problem definition into a hypothesis testing, we need to spell it out into the null (H_0) and alternative (H_a) hypotheses:

$$\begin{aligned} H_0 &: \overbrace{f(e1 \rightarrow e2 \mid \mathbf{X})}^{\substack{Posterior \\ \text{Model 1}}} > \overbrace{f(e2 \rightarrow e1 \mid \mathbf{X})}^{\substack{Posterior \\ \text{Model 2}}} \\ H_a &: \text{Otherwise} \end{aligned} \tag{4.1}$$

The null hypothesis[2] states that the density[3] of $f(e1 \rightarrow e2 \mid \mathbf{X})$ is centred towards a higher probability relative to $f(e2 \rightarrow e1 \mid \mathbf{X})$ (See Section 3.3.1 of how we conceptually compare two likelihoods and an exact idea can be applied to posteriors). In other words, we can conclude the direction as $e1 \rightarrow e2$ if H_0 holds. Otherwise, $e2 \rightarrow e1$. To put it another way, we wish to choose either Model 1 or Model 2. In hypothesis testing, the common scenarios involving H_0 and H_a, where you either reject H_0 in favour of H_a or fail to reject H_0. However, our approach allows for a third scenario, which we refer to as *uncertain* or *neither*. It occurs when the evidence gathered from the data is not strong enough to confidently support either H_0 or H_a hypothesis. In our study, the predicted direction can either be $e1 \rightarrow e2$, $e2 \rightarrow e1$ or *neither* of these. When we look at the models closely, they are

[2]Either '$>$' or '$\geq$' is acceptable since Model 1 and Model 2 are almost never computationally precisely equal to each other. Furthermore, we will later demonstrate how Bayes Factor manages if both models are indeed equal to each other.

[3]The highest probability density is called *mode* in statistics. When a distribution exhibits multimodal, comparing the two models this way may not be desirable. We explain an intuition here but will later demonstrate how to compare the models rigorously.

effectively the posteriors. Therefore, using the Bayes' rule as in Equation (3.4), we re-write the null hypothesis (4.1) as follows:

$$H_0 : \overbrace{f(\mathbf{X} \mid e1 \to e2)}^{Likelihood}\overbrace{f(e1 \to e2)}^{Prior} > \overbrace{f(\mathbf{X} \mid e2 \to e1)}^{Likelihood}\overbrace{f(e2 \to e1)}^{Prior} \tag{4.2}$$

It is worth noting that the denominator, $f(\mathbf{X})$, can be omitted because this term appears in both sides of the equation (See Equation (3.5) for more information). In what follows, we will explore likelihoods, priors and posteriors individually, particularly how we compute them.

Comments

By applying Bayes' Rule as in Equation (4.2), the posterior can be decomposed into the likelihood $f(\mathbf{X} \mid e2 \to e1)$, and prior $f(e2 \to e1)$. Consequently, for the computation of $f(e1 \to e2 \mid \mathbf{X})$ and $f(e2 \to e1 \mid \mathbf{X})$, we employ the Monte Carlo Markov Chain method (Lambert, 2018). It is worth noting that Stan takes care the numerical computations involved in this process (See Section 3.3.2). Lopez-Paz and Oquab (2018) is about a two-sample test when we have some samples and want to determine if they have significantly different distributions. In other words, if we collect samples from $f(e1 \to e2 \mid \mathbf{X})$ and samples from $f(e2 \to e1 \mid \mathbf{X})$ and then use a two-sample test to compare these samples, we are essentially testing whether the samples come from the same underlying distribution. This test can determine whether the two distributions are significantly different, but it does not directly address which causal direction, $e1 \to e2$ or $e2 \to e1$, should be concluded.

4.2.2 Likelihoods

Given $e1$ and $e2$ from an instance in the test set, we want to determine how likely we encounter the training set if $e1 \to e2$ holds, similarly if $e2 \to e1$ holds. Indeed, the likelihoods, $f(\mathbf{X} \mid e1 \to e2)$ and $f(\mathbf{X} \mid e2 \to e1)$, are exactly what we are after. If the training set supports $e1 \to e2$, we anticipate[4]

$$\mathbb{E}[f(\mathbf{X} \mid e1 \to e2)] = 1 \tag{4.3}$$

and

$$\mathbb{E}[f(\mathbf{X} \mid e2 \to e1)] = 0 \tag{4.4}$$

That is, the mean of $f(\mathbf{X} \mid e1 \to e2)$ centres largely around 1 whereas the mean of $f(\mathbf{X} \mid e2 \to e1)$ centres predominantly around 0. Let us see why it is the case. As we have two possible directions, $e1 \to e2$ or $e2 \to e1$, the sum of Equation (4.3) and (4.4) must be satisfied by Equation (4.5):

$$\mathbb{E}[f(\mathbf{X} \mid e1 \to e2)] + \mathbb{E}[f(\mathbf{X} \mid e2 \to e1)] = 1 \tag{4.5}$$

[4]The mean is identical to the mode when the distribution is symmetric unimodal. In fact, the likelihoods are always the case in our problem set-up. Equation (4.5) and (4.6) show contradiction if the distribution is multimodal.

Furthermore, if $\mathbb{E}[f(\mathbf{X} \mid e1 \rightarrow e2)] = \beta$, where $0 < \beta < 1$, then

$$\mathbb{E}[f(\mathbf{X} \mid e2 \rightarrow e1)] = 1 - \beta \tag{4.6}$$

in order to satisfy Equation (4.5). As both $\mathbb{E}[f(\mathbf{X} \mid e1 \rightarrow e2)] > 0$ and $\mathbb{E}[f(\mathbf{X} \mid e2 \rightarrow e1)] > 0$, they imply the training set supported $e1 \rightarrow e2$ to some extent, but the same training set would also support $e2 \rightarrow e1$. We have established bidirectional causal relations are not allowed (See Section 3.2). Hence, only the following conditions must hold:

$$\mathbb{E}[f(\mathbf{X} \mid e2 \rightarrow e1)] = \begin{cases} 0, & \text{if } \mathbb{E}[f(\mathbf{X} \mid e1 \rightarrow e2)] = 1 \\ 1, & \text{if } \mathbb{E}[f(\mathbf{X} \mid e1 \rightarrow e2)] = 0 \end{cases}$$

Unfortunately, when $\mathbb{E}[f(\mathbf{X} \mid e1 \rightarrow e2)]$ or $\mathbb{E}[f(\mathbf{X} \mid e1 \rightarrow e2)]$ is equal to 0 or 1 precisely, this will impose a computational challenge because given the likelihoods are absolutely certain, the priors will not be able to contribute at all. While it is logically valid, we need to dampen an effect such that the priors can still influence. Hence, we introduce a random variable called ε, which follows an uniform distribution (See Equation (3.1)):

$$\varepsilon \sim \mathit{Uniform}(0, 0.01) \tag{4.7}$$

What we have defined Equation (4.3) and (4.4) now become

$$\mathbb{E}[f(\mathbf{X} \mid e1 \rightarrow e2)] = 1 - \mathbb{E}[\varepsilon] = 0.995 \tag{4.8}$$

and

$$\mathbb{E}[f(\mathbf{X} \mid e2 \rightarrow e1)] = \mathbb{E}[\varepsilon] = 0.005 \tag{4.9}$$

respectively[5]. Effectively, we imply

$$f(\mathbf{X} \mid e1 \rightarrow e2) \sim \mathit{Uniform}(0.99, 1) \tag{4.10}$$

and

$$f(\mathbf{X} \mid e2 \rightarrow e1) \sim \mathit{Uniform}(0, 0.01) \tag{4.11}$$

respectively. As the mean is neither 0 nor 1, the priors can now contribute. Since the dataset is small, in most cases where both $e1$ and $e2$ cannot be found in the training set, the likelihoods do not favour one or other. Therefore, to compute the posteriors, we need to turn to the priors, $f(e1 \rightarrow e2)$ and $f(e2 \rightarrow e1)$, which we will look into word occurrences from the Internet's domains.

[5]Since ε follows an uniform distribution, we can always compute the mean (See Equation 3.2). Therefore, if $f(\mathbf{X} \mid e1 \rightarrow e2) \sim \mathit{Uniform}(0.99, 1)$, then $\mathbb{E}[f(\mathbf{X} \mid e1 \rightarrow e2)] = 0.995$.

Comments

Likelihoods are interpreted slightly differently depending on frequentist or Bayesian statistics. In frequentist statistics, the likelihood is a function of the parameters of the statistical model given the training data. It represents how well the data fits different values of the parameters, but it does not have any probability interpretation. However, in Bayesian statistics, the likelihood function represents the probability density of the training data given random variables (or parameters if you are a frequentist) (Lambert, 2018). In this context, random variables and the training data can have probability distributions. Since the likelihood function represents the probability density, it can have various distributions, including the uniform distribution.

In our study, the likelihood, denoted as $f(\mathbf{X} \mid e1 \rightarrow e2)$, provides insight into the probability distribution of encountering the training data when the causal relationship $e1 \rightarrow e2$ holds. Essentially, it captures all instances in the training data where $e1 \rightarrow e2$ is true. If $\mathbb{E}[f(\mathbf{X} \mid e1 \rightarrow e2)] = \gamma$, where γ is a fairly small scalar value, it indicates that some training data would support this argument, but the majority would not. In other words, it implies both $e1 \rightarrow e2$ and $e2 \rightarrow e1$ are observed in the training data to the different extents. This violates the underlying assumption that bidirectional causal relations are not allowed (See Section 1.3). Equation (4.3) is defined the expectation of *Uniform*(0.99, 1) in Equation (4.10).

4.2.3 Priors

In general, the priors can be anything that describes knowledge of $f(e1 \rightarrow e2)$ and $f(e2 \rightarrow e1)$. In our case, word occurrences from the Internet's domains that are used as a proxy to model causal directions become the priors[6]. The Internet has many domains, but not many would fit for the purpose of this project. Ideally, ones that fit should resemble closely to the dataset. To compute $P(e1 \rightarrow e2)$, which is a single probability, we count an occurrence of both $e1$ and $e2$ appear in a domain, $C(e1, e2)$, which is divided by an occurrence of $e1$ alone in the same domain, $C(e1)$. $P(e2 \rightarrow e1)$ can be calculated similarly. This will result in unnormalised versions, which will be normalised as described below.

$$P'(e1 \rightarrow e2) = \frac{C(e1, e2)}{C(e1)} \tag{4.12}$$

$$P'(e2 \rightarrow e1) = \frac{C(e1, e2)}{C(e2)} \tag{4.13}$$

In (4.12) and (4.13), $C(e1) \neq 0$ and $C(e2) \neq 0$ to avoid zero counts[7]. To normalise Equations (4.12) and (4.13), both are divided by their sum.[8]

[6] Word occurrences do not imply causal directions, but we use them as a surrogate.

[7] Haldane (1956) suggests adding 0.5 to every count if $C(e1) = 0$ or $C(e2) = 0$. However, we did not experience zero counts during the experiments.

[8] Bayesian statistics is inherently subjective in the sense that it allows individuals to express their beliefs through priors. Whether someone articulates $e1 \rightarrow e2$ or $e2 \rightarrow e1$ as expressed in Equations (4.14), (4.15), or any other forms, it remains an expression of their subjective belief.

$$P(e1 \to e2) = \frac{P'(e1 \to e2)}{P'(e1 \to e2) + P'(e2 \to e1)} \tag{4.14}$$

$$P(e2 \to e1) = \frac{P'(e2 \to e1)}{P'(e1 \to e2) + P'(e2 \to e1)} \tag{4.15}$$

Equations (4.14) and (4.15) are effectively conditional probabilities. We apply Equations (4.14) and (4.15) repeatedly for each domain outlined in Table 5.1. This process results in two distinct lists of probabilities, so we have $f(e1 \to e2)$ and $f(e2 \to e1)$. Now, we walk through one of the probability distributions, which can potentially be used to model the priors.

Uniform Distribution

If the prior, $f(e1 \to e2)$ or $f(e2 \to e1)$, follows an uniform distribution (See Section 3.1.1), we write

$$f(e1 \to e2) \sim \mathit{Uniform}(0, 1)$$

or

$$f(e2 \to e1) \sim \mathit{Uniform}(0, 1).$$

This distribution will be used if the likelihood $f(\mathbf{X} \mid e1 \to e2)$ or $f(\mathbf{X} \mid e2 \to e1)$ follows *Uniform*(0.99, 1) or *Uniform*(0, 0.01). That is, when the training set provides evidence, we will mostly rely on the likelihoods to tell us the causal direction.

Prior Specification

What happens if the likelihoods do not provide any evidence? In this case, we use probability distributions other than an uniform distribution to fit the priors. Prior specification is a process of selecting and defining a prior distribution in the Bayesian framework. More specifically, it involves choosing a type of distributions and its parameters. Given the experimental values computed from Equation (4.14), and (4.15) and a list of potential probability distributions as listed below:

- Cauchy distribution
- Exponential distribution
- Gamma distribution
- Inverse-gamma distribution
- Log-normal distribution
- Normal distribution (See Section 3.1.2)
- Student's *t*-distribution

Our goal is to find a distribution that fits the values best. How do we choose one? The criteria we have adopted is to minimise a Sum of Square Error (SSE). For a particular probability distribution of choice called j, we need to calculate SSE:

$$SSE_j = \sum_{i=1}^{N} (P_i - f_j^{-1})^2, \tag{4.16}$$

where i is the i^{th} Internet's domain, N is a total number of Internet's domains, P_i, which is short for $P(e1 \rightarrow e2)$ or $P(e2 \rightarrow e1)$, is the probability of the i^{th} domain computed from Equation (4.14), and (4.15), and f_j^{-1} is simulated samples from the probability distribution j. That is, if the probability distribution j fits the experimental values, simulated samples from the probability distribution j should look indistinguishable compared with the experimental values. Hence, we should have a small SSE. We repeat Equation (4.16) for all the potential probability distributions. The probability distribution that has the least SSE is deemed as the best distribution.

Prior Predictive Checks

SSE chooses what it is considered the best distribution, but how do we know if it is indeed the best fit? Kruschke (2015); Lambert (2018) suggest Prior Predictive Checks (PPC), which provides a guide to judge the fit. The concept is simple: if we cannot tell which data is generated from the probability distribution and which one comes from the experimental values as in Equation (4.14), and (4.15), we can conclude it is a good (enough) fit. Many statisticians use the maximum or minimum value as a criterion. That is, it is anticipated half of time the maximum or minimum value will come from simulated samples and another half it will come from experimental values if the chosen distribution fits the best. Algorithm 1 shows the pseudocode. M is a total number of runs that we ask the probability distribution to simulate samples; N is how many simulated samples we need for each run. Once N samples are generated, we retrieve the maximum or minimum value and store it in j. We also retrieve the maximum or minimum value from the experimental values and store it in k. If $j \geq k$ holds, we increment c by 1. Thus, c/M, which is the last line in Algorithm 1, is the percentage of times the maximum or minimum values come from simulated samples across M runs. If the probability distribution fits the best, we expect

$$c/M \approx 50\% \tag{4.17}$$

Algorithm 1 Prior Predictive Checks

Require: $m \geq 0, n \geq 0, i \geq 0$
  $M \leftarrow m$
  $c \leftarrow 0$
  while $M \neq 0$ **do**
    $N \leftarrow n$
    $i \leftarrow 0$
    while $N \neq 0$ **do**
      $p \leftarrow pdf(\theta)$
      $S[i] \leftarrow p$
      $i \leftarrow i + 1$
      $N \leftarrow N - 1$
    end while
    $j \leftarrow max(S)$ ▷ Or $min(S)$
    $k \leftarrow max(P)$ ▷ Or $min(P)$
    if $j \geq k$ **then**
      $c \leftarrow c + 1$
    end if
    $M \leftarrow M - 1$
  end while
  return c/M

4.2.4 Posteriors

The focus of Bayesian inference lies on the posteriors, $f(e1 \rightarrow e2 \mid \mathbf{X})$ and $f(e2 \rightarrow e1 \mid \mathbf{X})$. Using the priors and likelihoods, we can apply the Bayes' rule as in Equation (3.4) to compute the posteriors. Except for simple cases, estimating the posteriors will almost always be done computationally. Using Stan (See Section 3.3.2), we can compute the posteriors $f(e1 \rightarrow e2 \mid \mathbf{X})$ and $f(e2 \rightarrow e1 \mid \mathbf{X})$ easily.

4.2.5 Model Evaluation

Given both Model 1 and Model 2 are literally probability distributions, how do we compare them? It is tempting to use point estimators. That is, we could collapse the distribution into one single number using either the Posterior Mean, Posterior Median or Maximum A Posteriori (MAP) estimator. When the posterior exhibits a symmetric single-mode normally distributed curve, all three-point estimators will yield the same results. Often, all three almost always yield different results. A better way to handle this situation is Bayes Factor (BF) (Lambert, 2018; McElreath, 2015). While we are on the topic of model evaluation, it is worth to point out the differences between PPC and BF. PPC compares the simulated samples with the experimental values to determine if the chosen distribution fits the experimental values the best. However, BF is to compare weights between the two models to choose one over the another. While a model may have the most fitted distribution for the prior, as in PPC, it

may not be chosen when compared with another model due to BF. PPC may look redundant at first glance when we eventually use BF to compare the models. However, the job of PPC is to validate the prior distribution to make the analysis more statistical sound. Ultimately, the purpose is to choose either Model 1 or Model 2. That is, we need to compute the posterior odds between Model 1 and Model 2:

$$\frac{f(\text{Model 1} \mid \mathbf{X})}{f(\text{Model 2} \mid \mathbf{X})} \tag{4.18}$$

Using the Bayes' rule as in Equation (3.4), we rewrite the above equation as follows:

$$\frac{f(\mathbf{X} \mid \text{Model 1}) f(\text{Model 1})}{f(\mathbf{X} \mid \text{Model 2}) f(\text{Model 2})} \tag{4.19}$$

BF is simply defined as the ratio of two marginal likelihoods:

$$\text{BF} = \frac{f(\mathbf{X} \mid \text{Model 1})}{f(\mathbf{X} \mid \text{Model 2})} \tag{4.20}$$

If BF is larger than 1 (See Section 5.1.5 of how we manage when $BF = 1$), we can conclude Model 1. Otherwise, Model 2. As BF moves further away from 1, the more certain the model is. In fact, as it becomes closer to a positive infinity, Model 1 will definitely be our choice. Similarly, when it gets closer to 0, Model 2 will certainly be preferred. How we interpret BF can be found in Section 5.1.5. This concludes the first phase of the proposed method.

Comments

In Bayesian statistics, distinct tools and methodologies are applied for performing test statistics. In frequentist hypothesis testing, the comparison is typically made between a point estimate such as a sample mean and a distribution to determine how far the point estimate is from what would be expected under a specific null hypothesis. In other words, it assesses the degree of discrepancy between the observed data and the null hypothesis assumptions about the population parameter.

However, Bayesian hypothesis testing differs fundamentally from frequentist counterpart. It does not rely on the same concept of p-values or significant levels to perform hypothesis tests. Instead, Bayesian statistics focuses on calculating posterior distributions. Considering that Model 1 and Model 2 in the hypothesis test, as represented in Equation (4.1), are posterior distributions, we are effectively comparing two probability distributions. A commonly employed tool for this purpose is BF. BF serves as a measure to quantify the relative strength of evidence provided by the data in favour of one hypothesis, such as H_0, as opposed to another hypothesis, such as H_a.

4.3 BERT

While the Bayesian framework is capable to identify causal directions, a lack of understanding semantic means its capability is rather limited. Therefore, we turn into pre-trained language models. BERT (Devlin et al., 2019) is the language model used in NLP. Indeed, Tran Phu and Nguyen (2021); Zhao et al. (2021) use BERT to tackle causal relation identifications. Although BERT has many variants, we stick with a classical BERT — Bert Uncased Base Model. Our implementation is largely based on Rothman (2021)[9] and we summarise it into three major steps:

1. **Tokensization**: All sentences in the dataset have two entities marked by `<e1>` and `<e2>`, which do not have any semantic meanings to BERT. Hence, we first remove them and then convert the entire dataset into a CSV format, as shown in Table 4.3 for the first row, in order to facilitate text processing.

sentence-source	Label	Label-notes	Sentence
32	0	Comment:	He had chest pains and headaches from mold in the bedrooms.

Table 4.3 Sample data after pre-processing is shown.

   sentence-source is an identifier, which we do not process; Label is a target variable (i.e. 0 as $e1 \rightarrow e2$; 1 as $e2 \rightarrow e1$); Label-notes is a dummy; Sentence is an input sentence. We add two special tokens — `[CLS]` and `[SEP]` — to mark a beginning and end of the sentence. Additionally, BERT requires to set the maximum sequence length. We set it to 128 whereas the longest sequence in the dataset is 57.

2. **Classification**: Most neural network models require to set hyper-parameters initially, such as dropout rate, hidden size, etc. Instead, we use the ones provided by BERT. However, a few of them we still choose on our own. For instance, the batch size is 32; the number of epochs is set to 10.

3. **Evaluation**: The BERT model is primarily evaluated by F1 score.

In conclusion, although the Bayesian framework is capable of identifying causal directions, it is limited in its capability due to a lack of semantic understanding. To overcome this limitation, we leverage BERT. Our objective is to investigate whether the Bayesian framework can enhance BERT's performance. If successful, similar performance gains are likely to extend to other BERT variants. This indeed concludes the second phase.

[9] https://github.com/PacktPublishing/Transformers-for-Natural-Language-Processing/blob/main/Chapter02/BERT_Fine_Tuning_Sentence_Classification_DR.ipynb

4.4 Remarks

A traditional ensemble method involves combining outputs from multiple models to improve predictive performance. While the proposed method shares some similarities with the ensemble method, it exhibits distinctions in several key aspects:

- The Bayesian framework makes decisions based on confidence scores (i.e., Bayes Factor). As a result, the framework may decide not to make a decision if it lacks confidence. The traditional ensemble method involves weights of models' outputs, but not thresholds in decision making.
- The role of BERT is to refine the predictions from the Bayesian framework, rather than to combine predictions from the Bayesian framework.

In the case the proposed method was considered as an ensemble method, the removal of either the Bayesian framework or BERT would not render the method non-functional. In the unsupervised learning scenario, the absence of the Bayesian framework would disrupt the method's functionality since BERT lacked data from learning. Similarly, eliminating BERT would enable the Bayesian framework to make some decisions, but not all the decisions, given Bayes Factor. In data augmentation, removing the Bayesian framework, the proposed method would not impede the method's operation since BERT could still learn from the training data. However, the removal of BERT would enable the Bayesian framework to make some decisions, but not all the decisions, given Bayes Factor.

Chapter 5

Experiments

"The aim of a linguistic science is to be able to characterize and explain multitude of linguistic observations circling around us, in conversations, writing, and other media."

- Christopher D. Manning and Hinrich Schütze, *Foundations of Statistical Natural Language Processing (Manning and Schütze, 1999)*

In this chapter, we provide details of the experiments, walk through practical challenges, and finally present the experiment results. To recall earlier from Section 1.2, we have two research questions to answer:

1. **Data Augmentation**: *Does the Bayesian framework that we use to augment data boosts performance for BERT?* To answer this one, we have three experimental set-ups:
   (a) **Bayesian Framework**: In the first set-up, only the first phase is involved. The Bayesian model is trained on the training set and is tasked to predict on the test set. Basically, this set-up serves as a benchmark.
   (b) **BERT**: In the second set-up, only the second phase is involved. That is, BERT is trained on the training set and is tasked to predict on the test set. Like the previous one, this set-up also serves as a benchmark.
   (c) **Bayesian Framework + BERT**: In the final set-up, we employ the Bayesian model to enhance the training set before running BERT. The model is trained using both the training set and word occurrences to generate predictions for the test set. These predictions fall into one of the three categories: $e1 \rightarrow e2$, $e2 \rightarrow e1$ or *neither* based on Bayes Factors. When the predictions are either $e1 \rightarrow e2$ or $e2 \rightarrow e1$, corresponding test instances become as additional data for BERT. BERT in turn incorporates both the original training set and additional data created from the Bayesian framework to generate predictions for the remaining test data. We detail how it fares compared with the previous two set-ups.
2. **Unsupervised Learning**: *How well does the proposed method handle when no training is supplied?* In other to answer this one, we have three set-ups:

(a) **Random**: This set-up simply serves as a benchmark, which blindly guesses causal directions.

(b) **Bayesian Framework**: In the second set-up, we perform only the first phase. No training set is provided, and the Bayesian model takes word occurrences as inputs to make predictions. The goal is to assess the model's prediction performance based solely on the priors. This set-up also serves as a benchmark for comparison.

(c) **Bayesian Framework + BERT**: The last set-up involves executing both phases. We begin by using the Bayesian model to generate data using the priors. Next, we feed this generated data into BERT, which then makes predictions on the remaining test set. This set-up allows us to leverage the strengths of both the Bayesian model and BERT for improved prediction performance.

As we provided the methodology in Chapter 4, we now detail the nuts and bolts of the proposed method.

5.1 Bayesian Framework

What we want is to show empirically whether the priors somewhat resemble causal directions using the Bayesian framework. In this section, we fill in implementation details of the proposed method from Section 4.2. In what follows, we step through how the experiments are set up.

5.1.1 Hypothesis Testing

In the context of *research question 1 — data augmentation*, we utilize the null hypothesis (4.2), which is repeated as:

$$H_0 : \overbrace{f(\mathbf{X} \mid e1 \rightarrow e2)}^{Likelihood}\overbrace{f(e1 \rightarrow e2)}^{Prior} > \overbrace{f(\mathbf{X} \mid e2 \rightarrow e1)}^{Likelihood}\overbrace{f(e2 \rightarrow e1)}^{Prior} \tag{4.2}$$

However, in the context of *research question 2 — unsupervised learning* where no training set is available, we simplify the null hypothesis (4.2) as follows:

$$H_0 : \overbrace{f(e1 \rightarrow e2)}^{Prior} > \overbrace{f(e2 \rightarrow e1)}^{Prior} \tag{5.1}$$

Effectively, the posteriors are also the priors. To summarize, in data augmentation, we employ the null hypothesis (4.2), while in unsupervised learning, we rely on the null hypothesis (5.1).

5.1.2 Likelihoods

To recall what the likelihoods[1] are from Section 4.2.2, if $e1 \rightarrow e2$ holds in the training set, the same two entities, $e1$ and $e2$, in an opposite direction must be false. That is, if

[1] The likelihoods are omitted in unsupervised learning.

$$f(\mathbf{X} \mid e1 \rightarrow e2) \sim Uniform(0.99, 1) \tag{5.2}$$

then

$$f(\mathbf{X} \mid e2 \rightarrow e1) \sim Uniform(0, 0.01) \tag{5.3}$$

must hold (See Section 4.2.2). Fig. 5.1 shows what the likelihoods look like of one example, which has two entities – *suicide* and *death* (See the example from Section 4.1.2. In fact, this example is used for illustrative purposes throughout this chapter). In fact, the training set supports *suicide* $\rightarrow$ *death*.

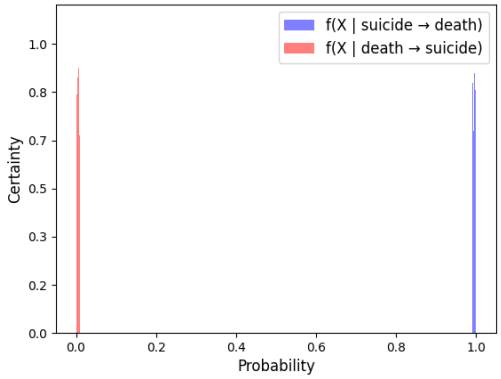

Fig. 5.1 Both likelihoods $f(\mathbf{X} \mid death \rightarrow suicide)$ and $f(\mathbf{X} \mid suicide \rightarrow death)$ are shown respectively.

5.1.3 Priors

As we explained in Section 4.2.3, to compute the priors, we count occurrences of $e1$ and $e2$ individually and both $e1$ and $e2$ together from the Internet's domains, and apply Equation (4.12) and (4.13). How do we get the occurrences in the first place? We make use of Google to look for the occurrences. For example, to look for the word *suicide* in the ABC News, the search command would be `suicide site:abc.net.au`. A number of search results, which we consider the occurrence, is shown before actual results are displayed in Fig. 5.2.

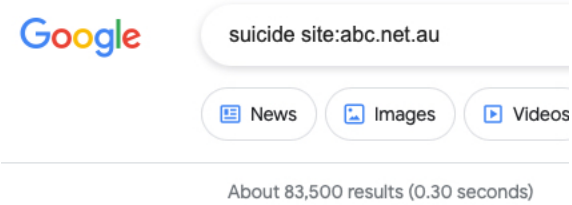

Fig. 5.2 When searching for the word *suicide*, 83,500 results are shown.

suicide appears 83,500 times whereas *death* appears 717,000 times in the ABC News, but only 47,000 times when both *suicide* and *death* are mentioned together. Hence, by employing the Equation (4.12) and (4.13), we have

$$P(suicide \rightarrow death) = \frac{C(suicide, death)}{C(suicide)} = \frac{47,000}{83,500} \tag{5.4}$$

and

$$P(death \rightarrow suicide) = \frac{C(suicide, death)}{C(death)} = \frac{47,000}{717,000} \tag{5.5}$$

respectively. We then apply normalisation as in Equation (4.14) and (4.15), so we end up

$$P(suicide \rightarrow death) = 0.8957 \tag{5.6}$$

and

$$P(death \rightarrow suicide) = 0.1043 \tag{5.7}$$

We repeat the same process for all other Internet's domains to get $f(suicide \rightarrow death)$ and $f(death \rightarrow suicide)$. Table 5.1 lists down all the Internet's domains we use to extract the word occurrences. We know SemEval-2010 (Task 8) is extracted from Wikipedia (See Table 2.1), which serves as a comprehensive resource for general knowledge in various areas such as science, news, arts and etc. The domains we employ in our study closely resemble these areas.

abc.net.au	au.news.yahoo.com	bbc.com
economist.com	edu	gov.au
imdb.com	mit.edu	nationalgeographic.com
ncbi.nlm.nih.gov	nejm.org	nytimes.com
oreilly.com	skynews.com.au	smh.com.au
springer.com	time.com	wikipedia.org
wiley.com		

Table 5.1 All the Internet's domains used for the priors are shown.

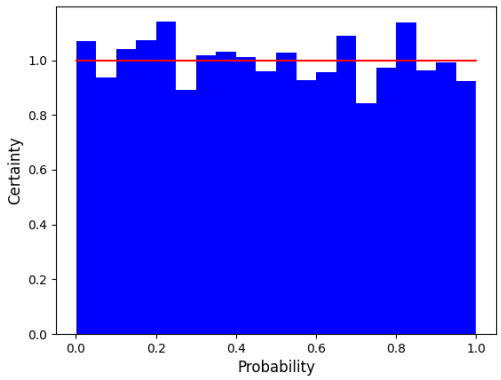

Fig. 5.3 Both priors $f(death \rightarrow suicide)$ and $f(suicide \rightarrow death)$ are equally likely.

Fortunately, we need not collect the occurrences in this case because the training set does provide evidence. Therefore, the priors follow an uniform distribution as shown in Fig. 5.3. Appendix B.1 shows the probabilities, $P(e1 \rightarrow e2)$ and $P(e2 \rightarrow e1)$, for all the Internet's domains in the test set.

Comments

In NLP, the order of concepts within a document or sentence holds significant importance. The order in which concepts appear in a document or sentence can affect the meaning or interpretation of the text. For instance, *John lifts an elephant up* and *an elephant lifts John up* are entirely different meanings. However, the prior, which is a probability distribution that represents beliefs or domain knowledge of a parameter, is often time independent. In general, a prior does not incorporate the order of concepts. However, in scenarios involving time-series data, the occurrence of a word at time t can be dependent on the words observed at time $t-1$. In such cases, Hidden Markov Model (HMM) (Manning and Schütze, 1999), which incorporates the order of concepts, can be the prior.

Prior Specification

Once we get the experimental values from Equation (4.14), and (4.15), we need to fit a distribution. As we previously established, the best distribution should have the smallest SSE. The `fitter` package[2] returns the best distribution based on the smallest SSE and parameters that describe the selected distribution. If we did not specify potential probability distributions, `fitter` would loop through 80 distributions[3]. Not only `fitter` takes longer to choose the best distribution, but also some of the distributions are not available in Stan, which means we cannot compute the posteriors. As one might speculate, a beta distribution would fit better as it is defined on the interval between 0 and 1. However, we found that its inclusion resulted in a deterioration of performance, leading us to remove it from consideration.

[2]https://fitter.readthedocs.io/en/latest/

[3]https://fitter.readthedocs.io/en/latest/tuto.html#histfit-class-fit-the-density-function-itself

Prior Predictive Checks

To validate the appropriateness of the selected distribution by `fitter`, we utilize PPC. PPC calculates the percentage of times the maximum or minimum value originates from simulated samples. Ideally, if the chosen distribution fits well, we expect this percentage to be close to 50% as indicated in Equation (4.17). However, requiring an exact 50% would be overly strict. Therefore, we have extended the range to 50% $\pm$ 1% to accommodate some variability. We investigate two different ranges: 50% $\pm$ 1% and 50% $\pm$ 50%. 50% $\pm$ 1% implies that between 49% and 51% of the time, the maximum or minimum value comes from simulated samples. Another range we consider is 50% $\pm$ 50%, which essentially means that we do not impose any checks. In our evaluation, we use the maximum value as a criterion for $f(e1 \rightarrow e2)$ and the minimum value for $f(e2 \rightarrow e1)$. The best range is determined by achieving the highest F1 score. However, we will present the results in Section 5.1.5 since we have not yet explained how to select between the two models and compute the F1 score.

5.1.4 Posteriors

Using the example from Section 4.1.2, Fig. 5.4 shows the posteriors after we multiply the likelihoods and priors together[4]. It shows $f(\textit{suicide} \rightarrow \textit{death} \mid \mathbf{X})$ is more likely since its density centres mostly around an upper end of probability relative to $f(\textit{death} \rightarrow \textit{suicide} \mid \mathbf{X})$. Visually, we know *suicide* is a cause of *death*, but we need to find a way to choose the models systematically.

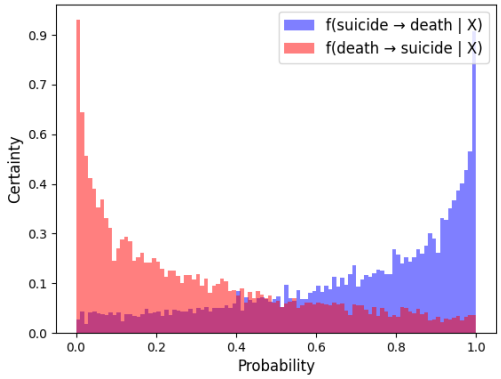

Fig. 5.4 Both posteriors $f(\textit{death} \rightarrow \textit{suicide} \mid \mathbf{X})$ and $f(\textit{suicide} \rightarrow \textit{death} \mid \mathbf{X})$ are shown.

5.1.5 Model Evaluation

To continue what we left off in Section 5.1.3, we need to determine the best range for PPC. Since we have two priors, $f(e1 \rightarrow e2)$ and $f(e2 \rightarrow e1)$, we would perform PPC separately for the same range. That is, one is for $f(e1 \rightarrow e2)$; another one is for $f(e2 \rightarrow e1)$. It is likely that two different best ranges could potentially result. To have the best range for both priors, we first perform PPC separately for

[4]In the unsupervised learning, the posterior distributions are also the priors distributions.

both $f(e1 \rightarrow e2)$ and $f(e2 \rightarrow e1)$ and BF. For each range, we then check whether PPC falls into the range (i.e. $50\% \pm 1\%$, $50\% \pm 2\%$, ..., $50\% \pm 50\%$) for both priors. If so, we consolidate them into one list and evaluate the results using the F1 score.

```
sentence-source = 9280  "The <e1>anxiety</e1> caused by the <e2>accident</e2
    >, which appears to show no sign of diminishing, and its negative impact
    on the living conditions in the affected areas, may be the principal
    reason for the increase in poor reported health."
Cause-Effect(e2,e1) Comment: modality is outside
```

Example 5.1 Sentence Source: 9280

While evaluating the models using BF, we have not yet taken uncertainty into account. That is, Equation (4.20) always chooses one model over the another. For instance, Example 5.1 is extracted from the test set. BF is computed as below:

$$\text{BF} = \overbrace{\underbrace{\frac{f(\mathbf{X} \mid \text{Model 1})}{f(\mathbf{X} \mid \text{Model 2})}}_{accident \rightarrow anxiety}}^{anxiety \rightarrow accident} = 0.16$$

In this case, BF chooses *accident* $\rightarrow$ *anxiety* despite either direction being almost equally likely. Worse still, the posteriors' plot as shown in Fig. 5.5 does not clearly distinguish them as a large portion of the overlapping area is seen. Hence, we must introduce a threshold. By enforcing the threshold, we allow the model to make predictions only if it is confident enough. Murphy (2013) provides a guideline of how to choose a model as shown in Table 5.2. In this project, we have simplified it to become Table 5.3 because we found the performance was somewhat similar among thresholds in Table 5.2. When BF lies on an extreme, either a positive infinity (in which case we consider as e^{300}) or close to 0 (in which case we consider as e^{-300}), it is very confident one model is preferred over the another.

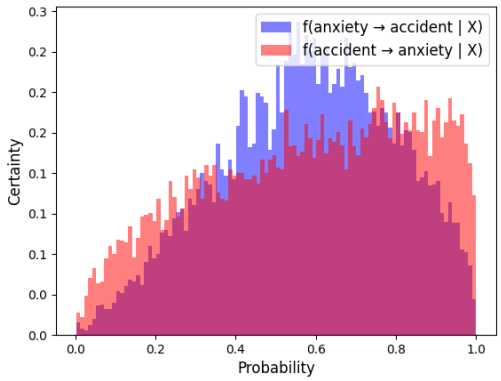

Fig. 5.5 Since the posteriors $f(accident \rightarrow anxiety \mid \mathbf{X})$ and $f(anxiety \rightarrow accident \mid \mathbf{X})$ heavily overlap, both *accident* $\rightarrow$ *anxiety* and *anxiety* $\rightarrow$ *accident* are likely.

BF	Interpretation
$BF < \frac{1}{100}$	Decisive evidence for Model 2
$BF < \frac{1}{10}$	Strong evidence for Model 2
$\frac{1}{10} < BF < \frac{1}{3}$	Moderate evidence for Model 2
$\frac{1}{3} < BF < 1$	Weak evidence for Model 2
$1 < BF < 3$	Weak evidence for Model 1
$3 < BF < 10$	Moderate evidence for Model 1
$BF > 10$	Strong evidence for Model 1
$BF > 100$	Decisive evidence for Model 1

Table 5.2 Jeffreys' thresholds of evidence for BF are shown.

BF	Interpretation
$BF < \mathrm{e}^{-300}$	Decisive evidence for Model 2
$\mathrm{e}^{-300} < BF < \mathrm{e}^{300}$	Uncertain or Neither Models
$BF > \mathrm{e}^{300}$	Decisive evidence for Model 1

Table 5.3 Thresholds that are used for the project are shown.

Reject Option

Uncertain can be interpreted "I don't know". Actually, the Bayesian model does know an answer, but it does not trust predictions much. In ML, this is called Reject Option (Bishop, 2007; Murphy, 2013). Bishop (2007) comments the reject region can be computed from the posteriors. Indeed, ours is $\mathrm{e}^{-300} < BF < \mathrm{e}^{300}$. In the experimental set-ups 1(c) and 2(c), where the Bayesian model generates data using the priors in which we feed this generated data into BERT, reject option tells us predicted directions that fall into reject option will *not* be fed into BERT. Now we introduce a cost function and see how we evaluate among Bayesian models with the different PPC ranges in the experimental set-up 1(a)(ii) and 2(b)(ii). Table 5.4 shows the cost of misclassification is λ_e. Clearly, to minimise the cost, one would choose not to predict at all, so we also introduce the cost λ_r should the model choose not to predict. For instance, if a model predicts incorrectly in 2 occasions out of 10, but it chooses not to predict in another 5 occasions, the total cost is $2\lambda_e + 5\lambda_r$. To determine λ_e and λ_r, we need to understand risk appetite. For instance, a ML application deployed in medical domains may be very risk averse. They would less likely tolerate a mistake but accept the application not to predict. In this case, λ_e is high, but low on λ_r.

		Estimate	
		$e1 \rightarrow e2$	$e2 \rightarrow e1$
Truth	$e1 \rightarrow e2$	0	λ_e
	$e2 \rightarrow e1$	λ_e	0

Table 5.4 Cost function that aims to evaluate among Bayesian models is shown.

5.2 BERT

As we can see, the Bayesian model can make predictions. However, it does so when it is confident enough. Hence, to classify the remaining samples, we employ BERT where we explained

BERT in Section 4.3. In the experimental set-up *unsupervised learning — Bayesian framework + BERT*, where we have a limited amount of labelled data from the Bayesian model, BERT can enhance the dataset through data augmentation techniques. One such approach involves utilizing `ContextualWordEmbsAug` from `nlpaug.augmenter` (Tunstall et al., 2022) to leverage contextual word embeddings.

```
from transformers import set_seed
import nlpaug.augmenter.word as naw

text = "Finally, Slone's fear of AIDS and the mental distress she suffered
    from this fear were caused by the needle stab."

aug = naw.ContextualWordEmbsAug(model_path = "bert-base-uncased", device = "
    cpu", action = "substitute")
print("Original: " + text)
print("Augmented: " + aug.augment(text)[0])
```

Example 5.2 Code using `ContextualWordEmbsAug`

Example 5.2 is the code snippet demonstrating `ContextualWordEmbsAug`. Example 5.3 shows the augmented text for one of the test instances. `action="substitute"` performs data augmentation through synonym replacement, where words with similar meanings are substituted within the text.

```
Original: Finally, Slone's fear of AIDS and the mental distress she suffered
    from this fear were caused by the needle stab.
Augmented: finally, slone's struggle with aids and the mental distress i
    suffered under this fear are caused from the biological error.
```

Example 5.3 Result after text is augmeneted

5.3 Results

Having walked through the details, we are now prepared to answer the research questions. All the experiments are conducted at the mention level. It is reminded that the F1 score is chosen to evaluate the model performance across all set-ups despite we also provide other metrics for reference. Each experimental set-up is run 10 times.

1(a) Data Augmentation — Bayesian Framework In this particular set-up, the Bayesian model underwent training using the provided training set and was then deployed to make predictions on the test data. We conducted two experiments to investigate different aspects. In the first experiment (i), we disabled the threshold in BF and turned off PPC. Essentially, this meant that the model was instructed (or rather compelled) to make predictions without considering its confidence level or the fulfilment of priors. In the second experiment (ii), we enabled the threshold in BF and selectively turned on or off PPC (See Section 5.1.3 for details). When PPC was activated with a range of 50% $\pm$

1%, it was considered *on*. Conversely, when the range was set to 50% ± 50%, PPC was considered *off*. In this case, the model would make predictions only if it exhibited sufficient confidence based on the established threshold.

Research Question	**Set-Up**	**Precision** (SD[b])	**Recall** (SD)	**F1** (SD)	**Accuracy** (SD)	**Coverage** (SD)	**Cost**[a] (SD)
1. Data Augmentation	(a)(i). Bayesian (BF−, PPC−)[c]	50.00% (0.00%)	56.72% (0.00%)	53.15% (0.00%)	59.15% (0.00%)	100.00% (0.00%)	– –
	(a)(ii). Bayesian (BF+, PPC+)	83.33% (0.00%)	100.00% (0.00%)	90.91% (0.00%)	94.29% (0.00%)	10.67% (0.00%)	$2\lambda_e + 293\lambda_r$ (0.00%)
	Bayesian (BF+, PPC−)	50.00% (0.00%)	57.52% (0.00%)	53.50% (0.00%)	60.35% (0.00%)	86.89% (0.00%)	$113\lambda_e + 43\lambda_r$ (0.00%)

Table 5.5 Results of experimental set-up 1(a) are displayed.

[a]Cost is introduced in Section 5.1.5. It is measured only when the model takes uncertainty into account (i.e., BF+).
[b]SD is short for Standard Deviation.
[c]"−" means *off* whereas "+" means *on*.

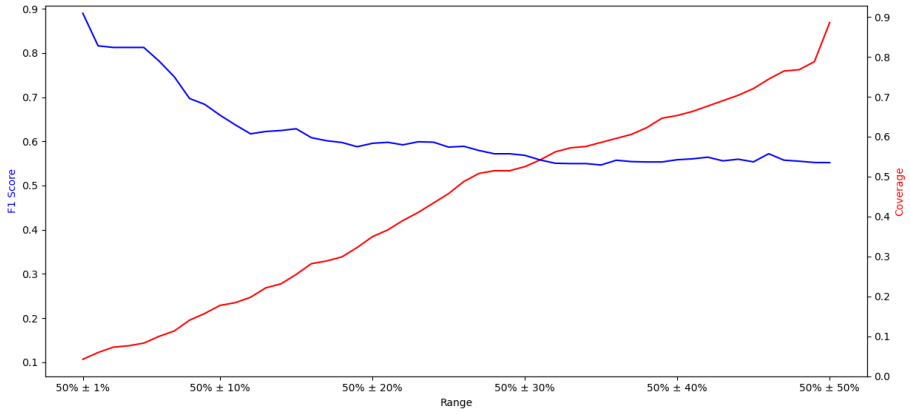

Fig. 5.6 How the performance changes while varying the PPC ranges.

Table 5.5 contains results for the first experiment (i), and you can refer to Table B.11, B.12, and B.13 for the specific individual results. Moving on to the second experiment (ii), its objective was to determine the optimal range for PPC. Fig.5.6 displays the F1 scores and coverage across various PPC ranges, with coverage representing the percentage of predictions made by the model. For a more detailed breakdown of the results, you can consult Table B.1, B.2, ..., B.10. Since the experiment was performed 10 times, we only present a single line for each metric in the figure because the results were consistently identical across all runs. Notably, when the range was set to 50% ± 1%, the model achieved an outstanding F1 score of 90.91%, which gradually decreased to 53.50% when using the wider range of 50% ± 50%. In contrast, the coverage exhibited a moderate increase from 10.67% at the range of 50% ± 1% to 86.89% at the range of 50% ± 50%.

The selection of models in (ii) is guided by the associated costs. However, since the values of λ_r and λ_e are unknown, Fig. 5.7 illustrates how the costs vary when these parameters are adjusted. In the figure, darker colors indicate higher costs, while lighter areas represent lower costs. Ideally, we aim to operate within the lighter regions to minimize the overall cost. If an individual exhibits a strong aversion to risk (i.e., high λ_e value), their preference would lean towards the (BF+, PPC+) model. Conversely, for those with a higher tolerance for risk (i.e., high λ_r value), the preference would be towards the (BF+, PPC−) model. The cost analysis assists in determining the most suitable model based on individual risk preferences.

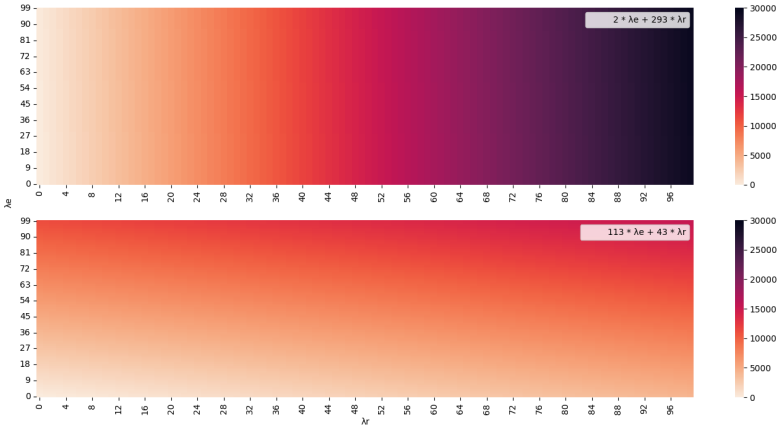

Fig. 5.7 The cost functions are presented based on whether PPC is enabled or disabled. The λ_r values are represented on the x-axis, while the y-axis corresponds to λ_e. In the top panel, the cost function $2\lambda_e + 293\lambda_r$ is displayed when PPC is enabled, while the bottom panel showcases the cost function $113\lambda_e + 43\lambda_r$ when PPC is disabled.

1(b) Data Augmentation — BERT In this set-up, only BERT was involved. That is, BERT was trained on the training set and tasked to predict on the test set. This experiment was run 10 times and averages were recorded in Table 5.6 (See Table B.14 for the individual runs).

Research Question	**Set-Up**	**Precision** (SD)	**Recall** (SD)	**F1** (SD)	**Accuracy** (SD)	**Coverage** (SD)	**Cost**
1. Data Augmentation	(b). BERT	95.06% (0.90%)	93.13% (1.37%)	94.08% (0.68%)	95.21% (0.53%)	100.00% (0.00%)	– –

Table 5.6 Results of experimental set-up 1(b) are displayed.

1(c) Data Augmentation — Bayesian Framework + BERT In this setting, we employed the Bayesian model to enhance the training set before running BERT. Put it simply, we use predictions

from 1(a)(ii) along with the training set as inputs to BERT. Results are shown in Table 5.7 (See Table B.15 and B.16 for the individual runs).

Research Question	Set-Up	Precision (SD)	Recall (SD)	F1 (SD)	Accuracy (SD)	Coverage (SD)	Cost
1. Data Augmentation	(c). Bayesian+BERT	94.94%	93.81%	94.34%	95.40%	100.00%	–
	(BF+, PPC+)	(2.08%)	(1.83%)	(0.66%)	(0.56%)	(0.00%)	–
	Bayesian+BERT	66.26%	76.94%	71.17%	74.54%	100.00%	–
	(BF+, PPC−)	(4.35%)	(5.19%)	(4.49%)	(3.91%)	(0.00%)	–

Table 5.7 Results of experimental set-up 1(c) are displayed.

In the case of Bayesian+BERT (BF+, PPC+), additional 15 instances were created. Conversely, in Bayesian+BERT (BF+, PPC−), a total of additional 286 instances were created.

2(a). Unsupervised Learning — Random This set-up purely serves as a benchmark. We simulated a probability from $Uniform(0,1)$. If it was greater than 0.5, we classified as $e1 \rightarrow e2$. Otherwise, $e2 \rightarrow e1$. We ran this set-up for 10,000 times and averages were recorded in Table 5.8.

Research Question	Set-Up	Precision (SD)	Recall (SD)	F1 (SD)	Accuracy (SD)	Coverage (SD)	Cost
2. Unsupervised Learning	(a). Random	40.81%	49.99%	44.90%	49.95%	100.00%	–
		(2.71%)	(4.32%)	(3.18%)	(2.74%)	(0.00%)	–

Table 5.8 Results of experimental set-up 2(a) are displayed.

2(b). Unsupervised Learning — Bayesian Framework In this particular set-up, we focused on conducting the first phase exclusively. The Bayesian model generated predictions solely based on the priors, without utilizing the training set.

Research Question	Set-Up	Precision (SD)	Recall (SD)	F1 (SD)	Accuracy (SD)	Coverage (SD)	Cost
2. Unsupervised Learning	(b)(i). Bayesian	46.00%	51.49%	48.59%	55.49%	100.00%	–
	(BF−, PPC−)	(0.00%)	(0.00%)	(0.00%)	(0.00%)	(0.00%)	–
	(b)(ii). Bayesian	0.00%	0.00%	0.00%	71.43%	2.13%	$2\lambda_e + 321\lambda_r$
	(BF+, PPC+)	(0.00%)	(0.00%)	(0.00%)	(0.00%)	(0.00%)	(0.00%)
	Bayesian	45.31%	51.33%	48.13%	55.36%	85.37%	$125\lambda_e + 48\lambda_r$
	(BF+, PPC−)	(0.00%)	(0.00%)	(0.00%)	(0.00%)	(0.00%)	(0.00%)

Table 5.9 Results of experimental set-up 2(b) are displayed.

Our experimentation involved two distinct scenarios. In the first experiment (i), we disabled the threshold in BF and excluded the usage of PPC. Essentially, this set-up compelled the model to

make predictions regardless of its level of confidence or the satisfaction of the priors. In the second experiment (ii), we enabled the threshold for BF and turned on or off for PPC (See Section 5.1.3). Under these conditions, the model only made predictions if it surpassed the confidence threshold and satisfied the priors according to PPC.

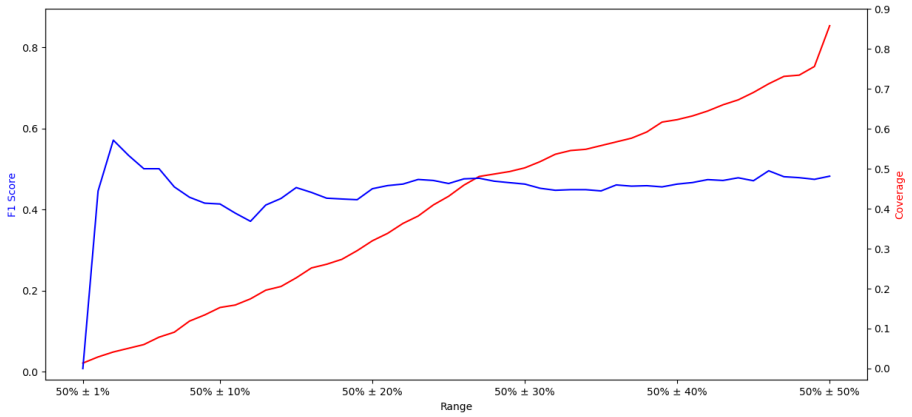

Fig. 5.8 How the performance changes while varying the PPC ranges.

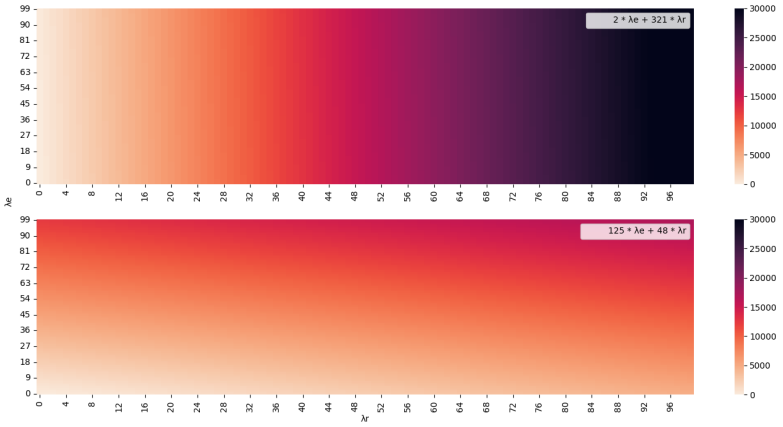

Fig. 5.9 The displayed cost functions depend on whether PPC is enabled or disabled. The x-axis represents λ_e, while the y-axis represents λ_r. The top panel illustrates the cost function is $2\lambda_e + 321\lambda_r$ when PPC is enabled, while the bottom panel shows the cost function $125\lambda_e + 48\lambda_r$ when PPC is disabled.

The results from the first experiment (i) have been recorded in Table 5.9 (See Table B.27, B.28, and B.29 for the detailed individual run data). In the second experiment (ii), we focused on determining the optimal range for PPC. Fig. 5.8 illustrates the F1 scores and coverages across various ranges. Since we conducted the experiment 10 times, only one line is displayed for each metric as the results were identical across all runs. The highest F1 score achieved was 57.14% at the range of $50\% \pm 3\%$

(See Table B.17), but it gradually decreased to 36.84% at the range of 50% ± 12% and remained relatively stable thereafter (See Table B.17, B.18, ..., B.26 for specific outcomes). Conversely, the coverage steadily increased from 2.13% at the range of 50% ± 1% to 85.37% at the range of 50% ± 50%. Fig. 5.9 depicts the variation of costs with respect to different values of λ_e and λ_r. It is evident that individuals with a strong aversion to risk (i.e., high values of λ_e) would favor (BF+, PPC+). Conversely, those who are more tolerant of risk (i.e., high values of λ_r) would lean towards the option of (BF+, PPC−).

2(c). Unsupervised Learning — Bayesian Framework + BERT In this setting, we executed both phases. Using those predictions from 2(b)(ii) as inputs to BERT, let us see how BERT performed for the remaining instances in the test set (See Table B.30 and B.31 for the individual runs). For (BF+, PPC+), we also performed BERT data augmentation (See Section 5.2) before running BERT for classification. Table 5.10 shows the results for all the experiments.

Research Question	Set-Up	Precision (SD)	Recall (SD)	F1 (SD)	Accuracy (SD)	Coverage (SD)	Cost
2. Unsupervised Learning	(c). Bayesian+BERT	46.00%	44.93%	44.89%	55.98%	100.00%	–
	(BF+, PPC+)	(2.56%)	(10.83%)	(6.22%)	(2.18%)	(0.00%)	–
	Bayesian+BERT	46.82%	52.09%	49.10%	56.31%	100.00%	–
	(BF+, PPC−)	(1.85%)	(7.86%)	(4.14%)	(1.39%)	(0.00%)	–

Table 5.10 Results of experimental set-up 2(c) are displayed.

5.3.1 Summary

Table 5.11 summarises the results of all experimental set-ups. Now, we want to find out if we have answered the research questions:

1. **Data Augmentation**: In the first set-up 1(a), when the Bayesian model incorporates uncertainty and performs PPC, it achieves significantly better results compared to the model without any checks. Specifically, it achieves a 90.91% F1 score for (BF+, PPC+), whereas the model without checks achieves a score of 53.15% for (BF−, PPC−), albeit at the expense of lower coverage. When we consider both set-ups 1(a) and 1(b) together, it becomes evident that BERT outperforms the Bayesian model across all the experiments.

   In comparing 1(b) and 1(c), the utilization of the Bayesian framework to augment data does not yield any noticeable improvement in BERT's performance. Specifically, the Bayesian+BERT model achieves an F1 score of 94.34% for (BF+, PPC+), whereas BERT alone achieves 94.08%. To determine whether the means of the F1 scores for the two models are truly equal, we conduct a two-sample *t*-test (Wackerly et al., 2002).

$$\begin{aligned} H_0 &: \mu_{(\text{BF+, PPC+})} - \mu_{\text{BERT}} = 0 \\ H_a &: \text{Otherwise} \end{aligned} \tag{5.8}$$

where $\mu_{(\text{BF+, PPC+})}$ is the mean F1 score in a population for (BF+, PPC+) and μ_{BERT} is the mean F1 score in a population for BERT. The t-statistics can therefore be calculated using Equation (5.9):

$$T = \frac{\bar{X}_{(\text{BF+, PPC+})} - \bar{X}_{\text{BERT}}}{\sqrt{\frac{s^2_{(\text{BF+, PPC+})}}{n_{(\text{BF+, PPC+})}} + \frac{s^2_{\text{BERT}}}{n_{\text{BERT}}}}} \tag{5.9}$$

where $\bar{X}_{(\text{BF+, PPC+})}$ is the sample average for (BF+, PPC+) and $\bar{X}_{\text{BERT}}$ is the counterpart for BERT; $s_{(\text{BF+, PPC+})}$ is the sample standard derivation for (BF+, PPC+) and s_{BERT} is the counterpart for BERT; $n_{(\text{BF+, PPC+})}$ is the sample size for (BF+, PPC+) and n_{BERT} is the counterpart for BERT. Let us compute Equation (5.9).

$$T = \frac{94.34 - 94.08}{\sqrt{\frac{0.66^2}{10} + \frac{0.68^2}{10}}} = 1.225 \tag{5.10}$$

At a significance level of 0.05, the t-value does not exceed the critical value of ± 2.101. Hence, we fail to reject the null hypothesis. The means of two models are not statistically significantly different at a significance level 0.05.

2. **Unsupervised Learning**: In the second set-up 2(b), when the Bayesian model incorporates uncertainty and regardless of PPC, it performs worse compared to the model without any checks. Specifically, it achieves a 48.59% F1 score (BF−, PPC−), while the model with (BF+, PPC+) achieves 0.00% and (BF+, PPC−) achieves 48.13%. However, if we consider accuracy as a metric, the additional gain is substantial, with a 71.43% accuracy for the model with (BF+, PPC+) compared to 55.49% for the model without any checks (BF−, PPC−).

   In the last set-up 2(c), when the Bayesian model generates data and feeds it into BERT, the Bayesian+BERT model achieves a slightly higher F1 score compared to the Bayesian model. Specifically, the Bayesian+BERT model achieves a 49.10% F1 score for (BF+, PPC–) while the Bayesian model achieves 48.59% for (BF–, PPC–). However, when we compare these results to those from set-up 2(a), the Bayesian+BERT model significantly outperforms random guessing, with a 49.10% F1 score compared to 44.90% for random guessing. Therefore, we conclude that the proposed method performs significantly better than random guessing when no training data is supplied.

In data augmentation, the Bayesian+BERT model performs best when PPC is activated whereas in unsupervised learning, the Bayesian+BERT model performs best when PPC is off. By comparing 2(a) with 2(c), we have shown empirically that word occurrences resemble the characteristics of causal directions, a 49.10% F1 score for Bayesian+BERT (BF+, PPC−) *vs* 44.90% for Random. We will explain the results further in Section 6.1.1.

Research Question	Set-Up	Precision (SD)	Recall (SD)	F1 (SD)	Accuracy (SD)	Coverage (SD)	Cost (SD)
1. Data Augmentation	(a)(i). Bayesian	50.00%	56.72%	53.15%	59.15%	100.00%	–
	(BF−, PPC−)	(0.00%)	(0.00%)	(0.00%)	(0.00%)	(0.00%)	–
	(a)(ii). Bayesian	83.33%	100.00%	90.91%	94.29%	10.67%	$2\lambda_e + 293\lambda_r$
	(BF+, PPC+)	(0.00%)	(0.00%)	(0.00%)	(0.00%)	(0.00%)	(0.00%)
	Bayesian	50.00%	57.52%	53.50%	60.35%	86.89%	$113\lambda_e + 43\lambda_r$
	(BF+, PPC−)	(0.00%)	(0.00%)	(0.00%)	(0.00%)	(0.00%)	(0.00%)
	(b). BERT	**95.06%**	93.13%	94.08%	95.21%	100.00%	–
		(0.90%)	(1.37%)	(0.68%)	(0.53%)	(0.00%)	–
	(c). Bayesian+BERT	94.94%	**93.81%**	**94.34%**	**95.40%**	100.00%	–
	(BF+, PPC+)	(2.08%)	(1.83%)	(0.66%)	(0.56%)	(0.00%)	–
	Bayesian+BERT	66.26%	76.94%	71.17%	74.54%	100.00%	–
	(BF+, PPC−)	(4.35%)	(5.19%)	(4.49%)	(3.91%)	(0.00%)	–
2. Unsupervised Learning	(a). Random	40.81%	49.99%	44.90%	49.95%	100.00%	–
		(2.71%)	(4.32%)	(3.18%)	(2.74%)	(0.00%)	–
	(b)(i). Bayesian	46.00%	51.49%	48.59%	55.49%	100.00%	–
	(BF−, PPC−)	(0.00%)	(0.00%)	(0.00%)	(0.00%)	(0.00%)	–
	(b)(ii). Bayesian	0.00%	0.00%	0.00%	71.43%	2.13%	$2\lambda_e + 321\lambda_r$
	(BF+, PPC+)	(0.00%)	(0.00%)	(0.00%)	(0.00%)	(0.00%)	(0.00%)
	Bayesian	45.31%	51.33%	48.13%	55.36%	85.37%	$125\lambda_e + 48\lambda_r$
	(BF+, PPC−)	(0.00%)	(0.00%)	(0.00%)	(0.00%)	(0.00%)	(0.00%)
	(c). Bayesian+BERT	46.00%	44.93%	44.89%	55.98%	100.00%	–
	(BF+, PPC+)	(2.56%)	(10.83%)	(6.22%)	(2.18%)	(0.00%)	–
	Bayesian+BERT	**46.82%**	**52.09%**	**49.10%**	**56.31%**	100.00%	–
	(BF+, PPC−)	(1.85%)	(7.86%)	(4.14%)	(1.39%)	(0.00%)	–

Table 5.11 All the experimental set-ups are displayed.

5.3.2 Remarks

In unsupervised learning, where we do not have the training set, the evaluation dataset is understandably unavailable. However, in the context of data augmentation, it is beneficial to split the training set, allocating a portion of it to serve as the evaluation dataset for model evaluation. We consider creation of the evaluation dataset is a valuable considerable in the data augmentation, which we intend to address in future work.

In the Bayesian framework, where the prior distribution varies for each test instance due to the entity pairs in the training set seldom matching those in the test set, the traditional k-fold cross validation may not be the most suitable approach to assess the goodness of fit for the prior distribution. Even if the model fits well on the evaluation set, it provides limited insight into how well the prior distribution fits the test set. Having said that, bootstrapping, which involves repeatedly drawing samples from the training dataset, could be used to evaluate the model's performance on these samples. Furthermore, in the unsupervised learning scenario, where we do not have the training dataset, the traditional cross validation is not even feasible.

Causal relations can change when observed datasets are changed. Ultimately, causal directions lie on Equation (4.2), where it consists of a likelihood and a prior. If the prior is uninformative, any changes to the training data, $\mathbf{X}$, can lead to changes in the posteriors, potentially leading to different causal directions. However, if the prior is strong, in other words the prior conveys strong belief or a high degree of confidence, any changes to the training data may have a moderate impact on the posterior or no impact at all.

5.4 Error Analysis

While a single metrics such as the F1 score may be the best way to evaluate ML models, Molnar (2022) argues it is an inadequate description of many tasks. Interpretability is rather more important, especially in risky domains. In this section, we delve deeper into the set-up 2(b)(ii) and aim to gain insights into why the Bayesian model made specific predictions, particularly the incorrect ones. From an interpretable machine learning perspective, the Bayesian model is considered intrinsic due to its simple structure. In the context of unsupervised learning, the posteriors effectively serve as the priors. Therefore, to comprehend the mechanics of the model, we focus our investigation on the priors. The errors observed in the model can be attributed to word occurrences. To shed light on this, we examine two specific cases (Example 5.4 and 5.5), which we have randomly picked. In these cases, the model made incorrect predictions despite BF and PPC being enforced (i.e., BF+, PPC+). By analyzing these cases, we aim to gain a better understanding of the factors contributing to the model's mispredictions.

Sentence Source: 8191 The correct answer of Example 5.4 is *rain* $\rightarrow$ *cancellation*, but the model misclassified as *cancellation* $\rightarrow$ *rain*.

```
1 sentence-source = 8191  "<e1>Rain</e1> caused <e2>cancellation</e2> of the
    event in 1877, so enforcement of the new law had to wait until 1878."
2 Cause-Effect(e1,e2) Comment:
```

Example 5.4 Sentence Source: 8191

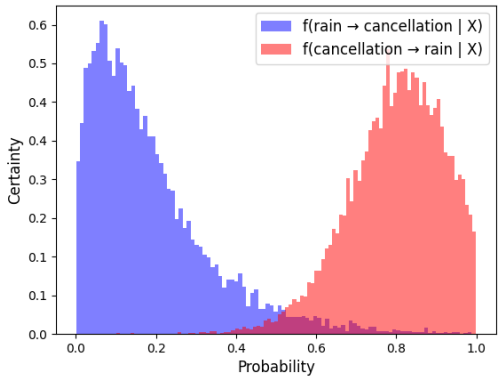

Fig. 5.10 Both $f(rain \rightarrow cancellation \mid \mathbf{X})$ and $f(cancellation \rightarrow rain \mid \mathbf{X})$ are shown. It is clear that the Bayesian model favours *cancellation* $\rightarrow$ *rain*.

Fig. 5.10, which shows the posteriors for both *rain* → *cancellation* and *cancellation* → *rain*, clearly favours $f(\textit{cancellation} \rightarrow \textit{rain} \mid \mathbf{X})$ because its density centres around an upper end of probability scale and a small overlapping area is seen. Since the priors were satisfied with PPC, we would not question the distributions that SSE chose. Instead, we look into the experimental values as shown in Table 5.12 (See Section 4.2.3). $P(\textit{cancellation} \rightarrow \textit{rain})$ across all the Internet's domains except au.news.yahoo.com are higher than $P(\textit{rain} \rightarrow \textit{cancellation})$. Based on Equation (4.12) or (4.13), when comparing the counts of the entities (i.e., $e1$ and $e2$), the entity with the higher count is more likely to be considered an effect. In this specific case, the count of *rain* is generally higher compared to *cancellation*. The reason *rain* appears more often in the text could be attributed to the fact that *rain* is commonly used in everyday language, particularly weather-related contexts like events related to weather conditions.

Causal Direction	wile y.com	gov. au	abc. net. au	natio nalge ograp hic .com	smh. com. au	oreill y.com	sprin ger .com	edu	wikip edia .org	nejm .org	bbc .com	time .com	mit .edu	au .news .yaho o.com	skyn ews .com .au	econo mist .com	ncbi .nlm .nih .gov	nyti mes .com	imdb .com
$P(\textit{rain} \rightarrow \textit{cancellation})$	0.03	0.28	0.04	0.07	0.18	0.32	0.05	0.26	0	0.18	0.07	0.43	0.17	0.54	0.3	0.21	0.17	0.15	0.07
$P(\textit{cancellation} \rightarrow \textit{rain})$	0.97	0.72	0.96	0.93	0.82	0.68	0.95	0.74	1.00	0.82	0.93	0.57	0.83	0.46	0.70	0.79	0.83	0.85	0.93

Table 5.12 Experimental values for sentence-source 8191 are shown.

Sentence Source: 8775 The correct answer of Example 5.5 is *moon* → *perturbations*, but the model misclassified as *perturbations* → *moon*.

```
sentence-source = 8775  "The thin F ring on the left of the image shows the <
    e1>perturbations</e1> caused by the <e2>moon</e2> Prometheus."
Cause-Effect(e2,e1) Comment:
```

Example 5.5 Sentence Source: 8775

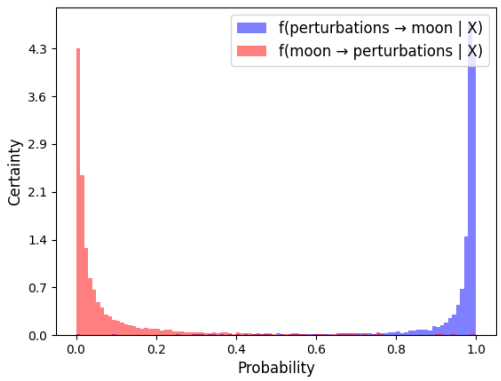

Fig. 5.11 Both $f(\textit{moon} \rightarrow \textit{perturbations} \mid \mathbf{X})$ and $f(\textit{perturbations} \rightarrow \textit{moon} \mid \mathbf{X})$ are shown. It is clear that the Bayesian model favours *perturbations* → *moon*.

Fig. 5.11, which shows the posteriors for both *moon* → *perturbations* and *perturbations* → *moon*, absolutely favours $f(\textit{perturbations} \rightarrow \textit{moon} \mid \mathbf{X})$ because its density centres around an upper end

of probability scale. Unlike the first case, density masses for both posteriors lie on an extreme. In other words, not much overlapping happens. Table 5.13 shows $P(perturbations \rightarrow moon)$ is higher than $P(moon \rightarrow perturbations)$ across all the Internet's domains except wiley.com, springer.com, and ncbi.nlm.nih.gov. However, one commonality among the three domains is, they provide access to scientific research articles, publications or resources. This case is more likely to be related to astronomy, so these specific domains more likely cover topics in this area.

Causal Direction	**wiley.com**	**gov.au**	**abc.net.au**	**nationalgeographic.com**	**smh.com.au**	**oreilly.com**	**springer.com**	**edu**	**wikipedia.org**	**nejm.org**	**bbc.com**	**time.com**	**mit.edu**	**au.news.yahoo.com**	**skynews.com.au**	**economist.com**	**ncbi.nlm.nih.gov**	**nytimes.com**	**imdb.com**
$P(perturbations \rightarrow moon)$	0.23	0.93	0.98	0.99	0.98	0.85	0.27	0.54	1	0.62	0.99	0.96	0.61	0.99	0.99	0.97	0.16	0.99	1
$P(moon \rightarrow perturbations)$	0.77	0.07	0.02	0.01	0.02	0.15	0.73	0.46	0.00	0.38	0.01	0.04	0.39	0.01	0.01	0.03	0.84	0.01	0.00

Table 5.13 Experimental values for sentence-source 8775 are shown.

In summary, the two aforementioned cases demonstrate that the Bayesian framework can fail not due to the methodology, but rather due to the specific word occurrences used to construct the priors. Having knowledge of the data would be beneficial in determining the appropriate domains to be utilized. For instance, in the second case where the sentence appears to be related to astronomy, domains like scientificamerican.com would be more suitable.

Chapter 6

Conclusion

> *"The answer to the question "why study causation?" is almost as immediate as the answer to "why study statistics." We study causation because we need to make sense of data, to guide actions and policies, and to learn from our success and failures."*
>
> - Judea Pearl, *Causal Inference in Statistics: A Primer (Pearl, 2016)*

Throughout the project, we have shown how the proposed method tackles causal direction identification. In this last chapter, we look at challenges as well as improvements we can make to overcome them.

6.1 Discussion

In this section, we analyse and interpret the results of the experiments. We also discuss limitations of the proposed method and explain some of the design choices.

6.1.1 Limitations

We discuss the limitations of the proposed method and list factors that might have influenced the results.

1. **Data Augmentation**: Let us revisit the set-up 1(a)(ii) in Section 5.3.1. When the Bayesian model took uncertainty into account and performed PPC as in (BF+, PPC+), it did noticeably better than the one without checks on PPC as in (BF+, PPC−). This can be explained from the fact that we had to discard a lot of priors as they failed PPC. Many priors chosen by SSE are not considered fit at all. That is, distributions listed in Section 4.2.3 do not in general capture the essential characteristics of word occurrences. We suspect this is due to two reasons:

   (a) We do not collect enough varieties of the Internet's domains. 8 out 19 domains belong to news whereas only 1 medicine domain is used. Google imposed a limit how many search

queries would be made programmatically within a certain time window, so it prevented us from collecting more observations.

(b) Distributions listed in Section 4.2.3 may not be right choices.

In comparing 1(b) and 1(c), the Bayesian+BERT model (BF+, PPC+) boosts the performance of BERT slightly further, but we should take it with a grain of salt as the difference is not statistically significant. We suspect BERT already had enough training data to learn from, so a couple of additional data points would not alter the result of BERT drastically.

2. **Unsupervised learning**: When we compare the best Bayesian+BERT models from each research question, the one from unsupervised learning has noticeably wider standard derivations across most metrics. Similar to the previous point, we are not sure about the suitability of the distributions listed in Section 4.2.3.

We will shortly look at Earth Mover's Distance (See Section 6.2.2) to see if we can better evaluate a chosen distribution. And we will also look at mixture models (See Section 6.2.3) to model the priors more properly.

6.1.2 Likelihoods Redefinition

The likelihoods[1] are almost impractical because of the data sparsity issue (See Section 5.1.2) . Put it simply, entities in both training and test sets do not have much in common. In fact, only 16 out of 328 instances the exact entity pairs from the test set can be found in the training set. We attempted to generalise the likelihoods by introducing a wildcard (i.e. $\cdot$), which means any word. Hence, Equation (5.2) and (5.3) became:

$$f(\mathbf{X} \mid e1 \rightarrow \cdot) \sim \mathit{Uniform}(0.99, 1) \tag{6.1}$$

and

$$f(\mathbf{X} \mid \cdot \rightarrow e1) \sim \mathit{Uniform}(0, 0.01) \tag{6.2}$$

respectively. Equation (6.1) can be interpreted as how likely the training set would be encountered if $e1$ was a cause of any noun. This modification would violate bidirectional causal directions. To know why it is the case, we shall first explain an intuition behind the redefinition. Example 6.1 is taken from the test set where the causal direction is *tsunami* $\rightarrow$ *death*, but we do not have the same pair in the training set. What we could do is to look up one of the two entities, say *death*, and work out if death is a cause or an effect from the training set. In most cases, we found death is more likely an effect such as *suicide* $\rightarrow$ *death*, *outbreak* $\rightarrow$ *death*, *infection* $\rightarrow$ *death* etc. Therefore, we would model $f(\mathbf{X} \mid \cdot \rightarrow death) \sim \mathit{Uniform}(0.99, 1)$, where $\cdot$ is *suicide*, *outbreak*, *infection* and etc. Nevertheless,

[1] The likelihoods are required in data augmentation, but not unsupervised learning.

we also found *death* being a cause in the training set such as $death \rightarrow breakdown$. In other words, we would also model $f(\mathbf{X} \mid death \rightarrow \cdot) \sim Uniform(0.99, 1)$. Therefore, we would create bidirectional causal directions, so we rolled back the implementation.

```
1 sentence-source = 8404  "The sheer scale of the <e1>death</e1> and
    destruction caused by the 2004 Boxing Day <e2>tsunami</e2> is impossible
    to fathom, even five years on."
2 Cause-Effect(e2,e1) Comment:
```

Example 6.1 sentence-source: 8404

One might argue to simply adjust parameters of the uniform distributions to reflect their relative likelihoods, but we previously established bidirectional causal directions would not be allowed in the training set. After all, we need two entities in order to establish causal relation. We want to emphasize the proposed method does not fail even if this assumption is violated because reject option from Section 5.1.5 could be interpreted as bidirectional causal directions.

6.2 Future Work

There are many areas we can do to improve or extend the project further. In this section, we share five of them: C4 Dataset, Earth Mover's Distance, Mixture Models, Bayesian Network and Counterfactual.

6.2.1 C4 Dataset

Rather than relying on Google for gathering word occurrences, C4 dataset[2] presents a viable alternative. This web crawl corpus consists of 15.7 million websites[3] from various domains, including news articles, blog posts, forum discussions, and etc. It offers several advantages:

- **Consistency**: It provides a static snapshot of web content, which allows us to work with a stable dataset, making it easier to replicate experimental results.
- **Preservation**: In contrast to dynamic websites that can change over time, C4 preserves data, which ensures valuable information is not lost.
- **Accessibility**: C4 is available offline, and it does not impose any limitations on the number of search queries can be made.

6.2.2 Earth Mover's Distance

When conducting PPC, we utilized the simple method to determine the percentage of times when the maximum or minimum value originated from simulated samples (See Section 4.2.3). This approach offers an advantage of being straightforward to implement because it involves comparing two numbers. However, it may not always provide reliable results. Gelman et al. (2004); Lambert (2018) recommend

[2]https://huggingface.co/datasets/c4

[3]https://searchengineland.com/search-websites-google-c4-dataset-395820

using the Kullback-Leibler Divergence (KL Divergence) to compare two distributions. However, the KL Divergence is sensitive to the choice of a reference distribution, which can be a drawback. An alternative measure for evaluating the best prior is the Earth Mover's Distance (EMD). Originally proposed in the field of image processing (Rubner et al., 11/2000), the same concept has been adopted in NLP as Word Mover's Distance (WMD) (Kusner et al., 2015; Sun et al., 2019). EMD is a methodology to compute a "distance" between the experimental values from Equation (4.14), and (4.15) and the listed distributions. The distribution with the shortest distance is considered as the best.

6.2.3 Mixture Models

Pre-selected distributions in Section 4.2.3 might not be flexible enough, so mixture models (Gelman et al., 2004) could be good substitutes. They are in fact probability distributions, which can account for data that exhibits multimodal and skewness. The idea is to take numerous probability distributions and stack them together using a linear combination. Let us revisit Section 3.1. If the light bulb market was flooded with light bulbs that had equally 20% as well as 30% chance of a defect, the probability distribution would look like Fig 6.1. To model the new probability distribution $g(D)$, we write as follows:

$$g(D) = \sum_{n=1}^{2} w_n f_n(D) \tag{6.3}$$

where w is a weight of each distribution. Since we have two underlying distributions, we weigh them accordingly to become the new probability distribution.

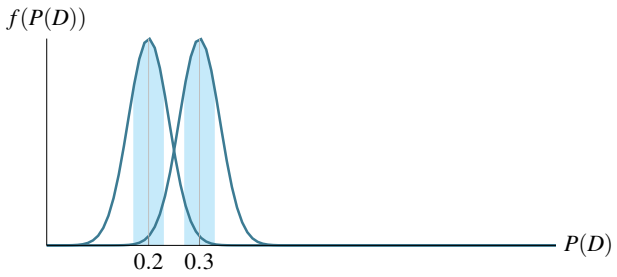

Fig. 6.1 If the light bulb market was flooded with light bulbs that had equally 20% as well as 30% chance of a defect, a probability distribution would look like multimodal.

6.2.4 Bayesian Network

We have so far considered a single causal relation in the sentence. To extend the analysis further, we may want to consider a multiple causal relations' scenario. That is, a model can determine the directions among all the causal relations. Pearl (2016) suggests Bayesian network may shed light on this problem. Indeed, Pearl (1988) proposes the Noisy-OR Model, where probabilities of a network

are satisfied with probability axioms and no directed cycles in the network are allowed. It would be better to illustrate the concept using an example. Let the following diagram, shown in the Fig. 6.2, be underlying causal relations. The task is to determine whether the causal direction $e3 \rightarrow e2$ exists. If $e3 \rightarrow e2$ exists, there are two possible networks as shown in Fig. 6.3 and 6.4.

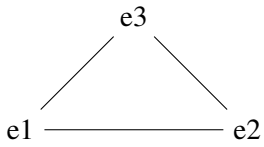

Fig. 6.2 $e1$, $e2$ and $e3$ show causal relations, but no directions are specified

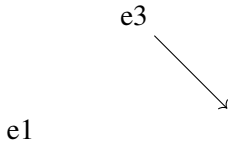

Fig. 6.3 $e3 \rightarrow e2$ is one possible way if $e3 \rightarrow e2$ exists.

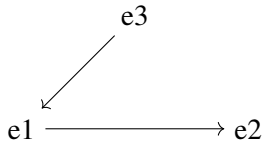

Fig. 6.4 $e3 \rightarrow e1 \rightarrow e2$ is another possible way if $e3 \rightarrow e2$ exists.

When datasets have multiple causal relations, Bayesian network, which is primarily based on the proposed method, can be constructed (Moghimifar et al., 2020).

6.2.5 Counterfactual

The work we have presented is the tip of the iceberg. Nevertheless, one area worthwhile to explore is called Counterfactual. In fact, we touched on counterfactual when we walked through SemEval-2020 (Task 5) in Section 2.2. To get the idea, let us visit the two hypothetical events that we previously walked through from the Section 2.3.4:

- $e1$: *A man speeded at 180km per hour in the Hume Highway.*
- $e2$: *Five people including a child injured in a head-on collision in the Hume Highway.*

A what-if question such as *what would an outcome be if the man did not speed?* would certainly pose a challenge to machines, but less so to human beings because human beings do not need data to infer a counter-outcome. Any ML models that rely on data would simply collapse to what-if questions because they do not have any data to infer from.

6.3 Final Thoughts

To revisit the two primary challenges outlined in Chapter 1, they are as follows: 1. Causal relations are few and far between in a document. 2. Implicit causal relations make identifying causal directions difficult. To tackle the first challenge, we have leveraged external sourced data to compensate for the scarcity of causal relations in the dataset. To address the second challenge, the Bayesian framework excels as it does not depend on explicit causal keywords like *cause*, making the predictions more robust. Through the experiments, we have shown word occurrences resemble the characteristics of causal directions. This finding provides valuable insights into the data generation process, thereby enhancing the ability of ML, including LLMs, to comprehend the relationships among entities in text. By incorporating prior knowledge into and understanding the Bayesian framework, we can enhance the quality and robustness of ML systems.

References

N. Asghar. Automatic extraction of causal relations from natural language texts: A comprehensive survey. *CoRR*, abs/1605.07895, 2016. URL http://arxiv.org/abs/1605.07895.

C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1 edition, 2007. ISBN 0387310738. URL http://www.amazon.com/Pattern-Recognition-Learning-Information-Statistics/dp/0387310738%3FSubscriptionId%3D13CT5CVB80YFWJEPWS02%26tag%3Dws%26linkCode%3Dxm2%26camp%3D2025%26creative%3D165953%26creativeASIN%3D0387310738.

L. Brinton and D. Brinton. *The Linguistic Structure of Modern English*. John Benjamins Publishing Company, 2010. ISBN 9789027211712. URL https://books.google.com.au/books?id=Q3uY9Ie0jWgC.

D.-S. Chang and K.-S. Choi. Causal relation extraction using cue phrase and lexical pair probabilities. In K.-Y. Su, J. Tsujii, J.-H. Lee, and O. Y. Kwong, editors, *Natural Language Processing – IJCNLP 2004*, pages 61–70, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg. ISBN 978-3-540-30211-7.

J. Chen, S. Lin, and G. Durrett. Multi-hop question answering via reasoning chains. *CoRR*, abs/1910.02610, 2019. URL http://arxiv.org/abs/1910.02610.

T. Dasgupta, R. Saha, L. Dey, and A. Naskar. Automatic extraction of causal relations from text using linguistically informed deep neural networks. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 306–316, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5035. URL https://aclanthology.org/W18-5035.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Q. Do, Y. S. Chan, and D. Roth. Minimally supervised event causality identification. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 294–303, 2011.

J. Dunietz, L. Levin, and J. Carbonell. Annotating causal language using corpus lexicography of constructions. In *Proceedings of the 9th Linguistic Annotation Workshop*, pages 188–196, Denver, Colorado, USA, June 2015. Association for Computational Linguistics. doi: 10.3115/v1/W15-1622. URL https://aclanthology.org/W15-1622.

L. Gao, P. K. Choubey, and R. Huang. Modeling document-level causal structures for event causal relation identification. In *Proceedings of the 2019 Conference of the North American Chapter*

of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 1808–1817, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1179. URL https://aclanthology.org/N19-1179.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition, 2004.

H. Gharagozlou, J. Mohammadzadeh, A. Bastanfard, and S. S. Ghidary. Semantic relation extraction: A review of approaches, datasets, and evaluation methods. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, apr 2023. ISSN 2375-4699. doi: 10.1145/3592601. URL https://doi.org/10.1145/3592601. Just Accepted.

R. Girju. Automatic detection of causal relations for question answering. In *Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering - Volume 12*, MultiSumQA '03, page 76–83, USA, 2003. Association for Computational Linguistics. doi: 10.3115/1119312.1119322. URL https://doi.org/10.3115/1119312.1119322.

R. Girju, P. Nakov, V. Nastase, S. Szpakowicz, P. Turney, and D. Yuret. SemEval-2007 task 04: Classification of semantic relations between nominals. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 13–18, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL https://aclanthology.org/S07-1003.

O. Glickman, I. Dagan, and M. Koppel. Web based probabilistic textual entailment. In *Proceedings of the 1st Pascal Challenge Workshop*, pages 33–36, 2005.

A. Gordon, Z. Kozareva, and M. Roemmele. SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 394–398, Montréal, Canada, 7-8 June 2012. Association for Computational Linguistics. URL https://aclanthology.org/S12-1052.

T. Griffiths and J. Tenenbaum. Theory-based causal induction. *Psychological Review*, 116(4): 661–716, 2009. URL http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=ovftk&NEWS=N&AN=00006832-200910000-00003.

B. J. B. S. Haldane. The estimation and significance of the logarithm of a ratio of frequencies. *Annals of Human Genetics*, 20, 1956.

I. Hendrickx, S. N. Kim, Z. Kozareva, P. Nakov, D. Ó Séaghdha, S. Padó, M. Pennacchiotti, L. Romano, and S. Szpakowicz. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL https://aclanthology.org/S10-1006.

C. Hidey and K. McKeown. Identifying causal relations using parallel Wikipedia articles. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1424–1433, Berlin, Germany, Aug. 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1135. URL https://aclanthology.org/P16-1135.

P. Hosseini, D. A. Broniatowski, and M. Diab. Predicting directionality in causal relations in text, 2021.

A. Ittoo and G. Bouma. Minimally-supervised learning of domain-specific causal relations using an open-domain corpus as knowledge base. *Data & Knowledge Engineering*, 88:142–163, 2013. ISSN 0169-023X. doi: https://doi.org/10.1016/j.datak.2013.08.004. URL https://www.sciencedirect.com/science/article/pii/S0169023X13000803.

D. Jurafsky and J. H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Pearson Education, Inc., Upper Saddle River, New Jersey, 2nd edition, August 2008. ISBN 9780131873216.

M. Kejriwal, C. A. Knoblock, and P. Szekely. *Knowledge Graphs: Fundamentals, Techniques, and Applications*. The MIT Press, Cambridge, MA, 2021.

V. Khetan, R. Ramnani, M. Anand, S. Sengupta, and A. E. Fano. Causal bert : Language models for causality detection between events expressed in text, 2021.

C. S. G. KHOO, J. KORNFILT, R. N. ODDY, and S. H. MYAENG. Automatic Extraction of Cause-Effect Information from Newspaper Text Without Knowledge-based Inferencing. *Literary and Linguistic Computing*, 13(4):177–186, 12 1998. ISSN 0268-1145. doi: 10.1093/llc/13.4.177. URL https://doi.org/10.1093/llc/13.4.177.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, Cambridge, Massachusetts, 1st edition, July 2009. ISBN 9780262013192.

P. Kroeger. *Analyzing Grammar: An Introduction*. Cambridge University Press, 2005. ISBN 9781139443517. URL https://books.google.com.au/books?id=rSglHbBaNyAC.

J. Kruschke. *Doing Bayesian Data Analysis (Second Edition)*. Academic Press, Boston, 2015.

M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger. From word embeddings to document distances. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 957–966, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/kusnerb15.html.

B. Lambert. *A Student's Guide to Bayesian Statistics*. SAGE Publications, 2018. ISBN 9781473916364. URL https://books.google.com.au/books?id=di9wswEACAAJ.

Z. Li, X. Ding, K. Liao, B. Qin, and T. Liu. Causalbert: Injecting causal knowledge into pre-trained models with minimal supervision, 2021.

D. Lopez-Paz and M. Oquab. Revisiting classifier two-sample tests, 2018.

K. Lu, I.-H. Hsu, W. Zhou, M. D. Ma, and M. Chen. Multi-hop evidence retrieval for cross-document relation extraction, 2022.

C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts, 1999. URL http://nlp.stanford.edu/fsnlp/.

D. Marcu and A. Echihabi. An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 368–375, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073145. URL https://aclanthology.org/P02-1047.

D. Mariko, H. A. Akl, E. Labidurie, S. Durfort, H. de Mazancourt, and M. El-Haj. Financial document causality detection shared task (fincausal 2020). *CoRR*, abs/2012.02505, 2020. URL https://arxiv.org/abs/2012.02505.

R. McElreath. *Statistical Rethinking, A Course in R and Stan*. Chapman and Hall/CRC, 2015. URL http://xcelab.net/rmpubs/rethinking/Statistical_Rethinking_sample.pdf.

F. Moghimifar, A. Rahimi, M. Baktashmotlagh, and X. Li. Learning causal bayesian networks from text, 2020.

C. Molnar. *Interpretable Machine Learning*. Lulu.com, 2 edition, 2022. URL https://christophm.github.io/interpretable-ml-book.

K. P. Murphy. *Machine learning : a probabilistic perspective*. MIT Press, Cambridge, Mass. [u.a.], 2013. ISBN 9780262018029 0262018020. URL https://www.amazon.com/Machine-Learning-Probabilistic-Perspective-Computation/dp/0262018020/ref=sr_1_2?ie=UTF8&qid=1336857747&sr=8-2.

J.-H. Oh, K. Torisawa, C. Hashimoto, T. Kawada, S. De Saeger, J. Kazama, and Y. Wang. Why question answering using sentiment analysis and word classes. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 368–378, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL https://aclanthology.org/D12-1034.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, 1st edition, September 1988. ISBN 9781558604797.

J. Pearl. *Causal inference in statistics : a primer*. Wiley, Chichester, West Sussex, 2016. ISBN 9781119186854.

J. Pearl and D. Mackenzie. *The Book of Why*. Basic Books, New York, 2018. ISBN 978-0-465-09760-9.

S. M. Reimann. Multilingual zero-shot and few-shot causality detection. Master's thesis, Uppsala University, Department of Linguistics and Philology, 2021.

D. Rothman. *Transformers for Natural Language Processing: Build Innovative Deep Neural Network Architectures for NLP with Python, PyTorch, TensorFlow, BERT, RoBERTa, and More*. Packt Publishing, 2021. ISBN 9781800565791. URL https://books.google.com.au/books?id=Ua03zgEACAAJ.

Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40, 11/2000. URL https://www.cs.cmu.edu/~efros/courses/LBMV07/Papers/rubner-jcviu-00.pdf.

M. Singer, M. Halldorson, J. C. Lear, and P. Andrusiak. Validation of causal bridging inferences in discourse understanding. *Journal of Memory and Language*, 31(4):507–524, 1992. ISSN 0749-596X. doi: https://doi.org/10.1016/0749-596X(92)90026-T. URL https://www.sciencedirect.com/science/article/pii/0749596X9290026T.

C. Sun, K. T. J. Ng, P. Henville, and R. Marchant. Hierarchical word mover distance for collaboration recommender system. In R. Islam, Y. S. Koh, Y. Zhao, G. Warwick, D. Stirling, C.-T. Li, and Z. Islam, editors, *Data Mining*, pages 289–302, Singapore, 2019. Springer Singapore. ISBN 978-981-13-6661-1.

P. Suppes. A probabilistic theory of causality. *British Journal for the Philosophy of Science*, 24(4): 409–410, 1973.

F. A. Tan, H. Hettiarachchi, A. Hürriyetoğlu, T. Caselli, O. Uca, F. F. Liza, and N. Oostdijk. Event causality identification with causal news corpus – shared task 3, case 2022, 2022.

Y. Tian, G. Chen, Y. Song, and X. Wan. Dependency-driven relation extraction with attentive graph convolutional networks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4458–4471, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.344. URL https://aclanthology.org/2021.acl-long.344.

M. Tran Phu and T. H. Nguyen. Graph convolutional networks for event causality identification with rich document-level structures. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3480–3490, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.273. URL https://aclanthology.org/2021.naacl-main.273.

L. Tunstall, L. von Werra, and T. Wolf. *Natural Language Processing with Transformers: Building Language Applications with Hugging Face*. O'Reilly Media, Incorporated, 2022. ISBN 9781098103248. URL https://books.google.com.au/books?id=_0uezgEACAAJ.

S. Verberne, L. Boves, N. Oostdijk, and P.-A. Coppen. Evaluating discourse-based answer extraction for why-question answering. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, page 735–736, New York, NY, USA, 2007. Association for Computing Machinery. ISBN 9781595935977. doi: 10.1145/1277741.1277883. URL https://doi.org/10.1145/1277741.1277883.

D. D. Wackerly, W. M. III, and R. L. Scheaffer. *Mathematical Statistics with Applications*. Duxbury Advanced Series, sixth edition edition, 2002.

J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le. Finetuned language models are zero-shot learners. *CoRR*, abs/2109.01652, 2021. URL https://arxiv.org/abs/2109.01652.

F. Xuelan and G. Kennedy. Expressing causation in written english. *RELC Journal*, 23(1):62–80, 1992. doi: 10.1177/003368829202300105. URL https://doi.org/10.1177/003368829202300105.

J. Yang, S. C. Han, and J. Poon. A survey on extraction of causal relations from natural language text. *CoRR*, abs/2101.06426, 2021. URL https://arxiv.org/abs/2101.06426.

X. Yang, S. Obadinma, H. Zhao, Q. Zhang, S. Matwin, and X. Zhu. SemEval-2020 task 5: Counterfactual recognition. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 322–335, Barcelona (online), Dec. 2020. International Committee for Computational Linguistics. doi: 10.18653/v1/2020.semeval-1.40. URL https://aclanthology.org/2020.semeval-1.40.

K. Zhao, D. Ji, F. He, Y. Liu, and Y. Ren. Document-level event causality identification via graph inference mechanism. *Information Sciences*, 561:115–129, 2021. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2021.01.078. URL https://www.sciencedirect.com/science/article/pii/S002002552100116X.

Appendix A

Methodology

A.1 Exploratory Analysis

We expand the comments mentioned in Section 4.1.2 to include samples from the test set. Specifically, we randomly examined 20 sentence-sources in the test set — 8005, 8027, 8031, 8041, 8043, 8055, 8058, 8073, 8083, 8105, 8107, 8108, 8116, 8118, 8159, 8175, 8191, 8204, 8219, and 8234. We raised concerns for two of them below:

sentence-source: 8041 The causal direction is *subject* → *implication*, but we are not sure if it is correct. We argue the two entities do not have any causal direction because *implication* is kind of a conclusion which one needs to derive from evidence. *subject* is simply the source of the implication.

```
1 sentence-source = 8041  "The <e1>subject</e1> of "imply" is the source of an
    <e2>implication</e2> while the subject of "infer" is the recipient of an
    implication."
2 Cause-Effect(e1,e2) Comment:
```

Example A.1 sentence-source: 8041

sentence-source: 8116 The causal direction is *laughing* → *joy*. Laughing is one of many activities Lisa did to contribute to joy, but we argue all the activities together that contributed to joy.

```
1 sentence-source = 8116  "Lisa took great <e1>joy</e1> from <e2>laughing</e2>,
     volunteering at school, taking pictures, chatting, the Twins and Vikings
    , playing softball and volleyball, and time at Lake Vermilion."
2 Cause-Effect(e2,e1) Comment:
```

Example A.2 sentence-source: 8116

Appendix B

Experiments

B.1 Priors

The tables below show the probabilities, $P(e1 \rightarrow e2)$ and $P(e2 \rightarrow e1)$, for all the Internet's domains in the test data. sentence-source is an identifier. All other fields are the Internet's domains. A blank cell implies no occurrences of $e1$ and $e2$ together.

B.1.1 $P(e1 \rightarrow e2)$

Sentence Source	wiley.com	gov.au	abc.net.au	nationalgeographic.com	smh.com.au	oreilly.com	springer.com	edu	wikipedia.org	nejm.org	bbc.com	time.com	mit.edu	au.news.yahoo.com	skynews.com.au	economist.com	ncbi.nlm.nih.gov	nytimes.com	imdb.com
8005	0.97	0.7	0.9	0.98	0.84	0.99	0.95	0.94	0.82	0.88	0.98	0.93	0.9	0.99	0.98	0.93	0.98	0.91	0.97
8027	0.53	0.68	0.87	0.8	0.84	0.76	0.4	0.65	0.93	0.49	0.95	0.84	0.66	0.41	0.83	0.57	0.35	0.95	0.76
8031	0.74	0.71	0.55	0.29	0.69	0.13	0.6	0.59	0.38	0.59	0.29	0.26	0.42	0.11	0.4	0.37	0.66	0.61	0.31
8041	0.7	0.22	0.29	0.22	0.24	0.28	0.65	0.24	0	0.46	0.14	0.07	0.32	0.22	0.21	0.32	0.48	0.2	0.05
8043	0.99	0.96	0.96	0.87	0.94	0.99	0.99	0.99	1	0.99	0.94	0.81	0.95	0.83	0.86	0.77	1	0.91	0.94
8055	0.64	0.66	0.18	0.17	0.43	0.6	0.63	0.59	0.01	0.59	0.25	0.39	0.38	0.53	0.23	0.55	0.59	0.23	0.03
8058	0.9	0.93	0.94	0.93	0.97	0.84	0.89	0.96	1	0.95	0.95	0.94	0.88	0.97	0.99	0.96	0.99	0.95	0.95
8073	0.03	0.74	0.75	0.52	0.63	0.18	0.08	0.21	0.94	0.04	0.9	0.45	0.47	0.49	0.83	0.45	0.01	0.48	1
8083	0.91	0.83	0.76	0.85	0.87	0.87	0.94	0.87	0.92	0.88	0.84	0.87	0.84	0.87	0.9	0.89	0.96	0.87	0.86
8105	0.68	0.17	0.79	0.52	0.9	0.5	0.41	0.4	0.78	0.5	0.5	0.52	0.59	0.42	0.63	0.49	0.33	0.61	0.53
8107	0.27	0.01	0.09	0.13	0.03	0.08	0.16	0.04	0	0.08	0.03	0.05	0.24	0.02	0.01	0.13	0.23	0.02	0.04
8108	0.01	0.01	0.02	0.02	0.03	0.01	0	0	0	0	0.02	0.02	0.05	0.14	0.01	0.05	0	0.06	0.22
8116	0.33	0.34	0.58	0.62	0.61	0.14	0.36	0.35	0.36	0.28	0.43	0.53	0.62	0.5	0.36	0.45	0.2	0.25	0.47
8118	0.6	0.41	0.73	0.84	0.88	0.48	0.61	0.7	0.61	0.85	0.85	0.84	0.58	0.7	0.54	0.91	0.44	0.61	0.91
8159	0.22	0.83	0.27	0.34	0.84	0.61	0.29	0.33	0.61	0.85	0.71	0.73	0.28	0.75	0.72	0.61	0.05	0.77	0.92
8175	0.67	0.84	0.96	0.49	0.87	0.87	0.53	0.9	1	0.85	0.92	0.66	0.81	0.72	0.84	0.77	0.34	0.86	1
8191	0.03	0.28	0.04	0.07	0.18	0.32	0.05	0.26	0	0.18	0.07	0.43	0.17	0.54	0.3	0.21	0.17	0.15	0.07
8204	0.01	0	0.01	0.06	0.01	0	0	0	0	0.01	0	0.01	0.01	0.01	0	0	0.01	0.01	0
8219	0.97	0.58	0.64	0.73	0.59	0.72	0.91	0.91	1	0.97	0.94	0.61	0.84	0.7	0.73	0.71	0.99	0.91	0.61
8234	0.01	0.07	0.03	0.03	0.09	0.19	0.01	0.09	0.04	0.14	0.04	0.07	0.11	0.03	0.04	0.09	0	0.08	0.08
8236	0.87	0.49	0.7	0.68	0.61	1	0.93	0.96	0.95	0.81	0.85	0.82	0.98	0.88	0.89	0.9	0.86	0.78	0.78
8239	0.1	0.12	0.4	0.01	0.19	0.03	0.07	0.32	0.1	0.48	0.24	0.46	0.24	0.18	0.21	0.31	0.07	0.43	0.31
8240	0.98	0.88	0.61	0.94	0.65	0.93	0.95	0.97	0.99	0.69	0.55	0.89	0.87	0.96	0.6	0.92	1	0.57	0.4
8253	0.99	0.87	0.87	0.97	0.92	1	0.98	0.99	0.89	1	0.8	0.8	0.97	0.72	0.65	0.8	1	0.75	0.97
8257	0.01	0.16	0.04	0.09	0.03	0.03	0.01	0.01	0.01	0	0.02	0.01	0.08	0.08	0.01	0.02	0	0.02	0
8265	0	0	0.04	0.16	0.03	0.08	0	0.01	0.05	0.01	0.08	0.11	0.06	0.11	0.09	0.03	0	0.3	0.22
8288	0.99	0.88	0.4	0.69	0.29	0.92	0.99	0.9	0.5	0.74	0.2	0.52	0.8	0.82	0.2	0.69	1	0.61	0.04
8312	0.2	0.11	0.16	0.03	0.02	0.04	0.24	0.02	0	0.07	0.03	0.07	0.14	0.05	0.01	0.24	0.11	0.04	0
8334	0.04	0.68	0.62	0.05	0.19	0.42	0.04	0.15	0.09	0.05	0.32	0.59	0.28	0.67	0.06	0.72	0.01	0.49	0.21
8357	0.24	0.65	0.91	0.63	0.29	0.64	0.2	0.35	0.75	0.6	0.77	0.89	0.5	0.8	0.75	0.81	0.25	0.78	0.82
8361	0.52	0.18	0.47	0.31	0.18	0.17	0.52	0.27	0.28	0.11	0.2	0.25	0.52	0.73	0.24	0.35	0.37	0.21	0.09
8373	0.01	0.02	0.04	0.16	0.17	0.47	0.01	0.06	0.12	0.07	0.17	0.27	0.19	0.02	0.13	0.18	0	0.25	0.6
8377	0.91	0.9	1		1		0.9	0.97	0.91	0.91	0.99	0.98	0.99	0.98	1	0.99	0.89	0.98	1
8382	0.02	0.02	0.1	0.25	0.08	0.13	0.02	0.08	0.09	0.03	0.09	0.22	0.16	0.02	0.14	0.19	0	0.15	0.26
8402	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1	1	0.97	0.49	0.98	0.98	1	0.99	0.98	1	0.99	1
8403	0.01	0.06	0.31	0.23	0.44	0.15	0.01	0.08	0.29	0.05	0.3	0.5	0.68	0.26	0.49	0.42	0	0.46	0.77
8404	0	0.03	0.02	0.05	0.03	0.09	0	0	0	0	0.01	0.04	0.06	0.1	0.02	0.1	0	0	0
8405	0.96	0.91	1	0.98	0.99	1	0.93	0.96	1	0.84	1	0.97	0.96	0.86	0.99	0.89	0.98	0.99	1
8409	0.97	0.99	0.99	0.99	0.99	0.97	0.98	1	1	0.96	1	1	0.93	0.99	1	0.98	0.99	1	1
8417	1	1	1		1	1	1	1	1	1	1	1	0.99	1			1	1	
8439	0.06	0.1	0.21	0.1	0.37	0.06	0.06	0.02	0.32	0.1	0.27	0.06	0.15	0.02	0.25	0.07	0.03	0.11	0.16
8455	0.01	0.05	0.08	0.04	0.04	0.1	0	0.15	0.49	0.01	0.37	0.11	0.08	0.13	0.1	0.11	0	0.15	0.05
8471	0.36	0.87	0.72	0.98	0.78	0.77	0.72	0.41	0.51	0.59	0.91	0.84	0.52	0.47	0.69	0.62	0.49	0.67	0.92
8473	0.81	0.88	0.48	0.43	0.76	0.61	0.84	0.78	0.56	0.86	0.42	0.5	0.55	0.29	0.49	0.42	0.96	0.68	0.31
8476	0.99	0.44	0.5	0.91	0.57	0.51	0.99	0.92	0.7	0.97	0.64	0.41	0.76	0.24	0.49	0.35	1	0.43	0.52
8486	0.99	0.76	0.87	0.91	0.82	0.69	0.98	0.86	1	0.97	0.93	0.79	0.74	0.86	0.75	0.52	1	0.96	0.97
8489	0.94	0.93	0.63	0.74	0.64	0.91	0.91	0.96	1	0.77	0.95	0.59	0.84	0.08	0.64	0.54	0.97	0.85	0.83
8493	0.05	0.01	0.19	0.11	0.03	0.16	0.16	0.04	0.04	0.02	0.03	0.05	0.24	0.01	0.03	0.06	0.06	0.04	0.02
8494	0.99	0.87	0.5	0.74	0.09	0.93	0.98	0.96	0.58	0.79	0.74	0.54	0.92	0.2	0.43	0.55	1	0.36	0.27
8498	0.33	0.4	0.35	0.22	0.49	0.58	0.43	0.39	0.28	0.45	0.53	0.53	0.33	0.63	0.62	0.53	0.39	0.52	0.46
8500	0.02	0.45	0.18	0.22	0.47	0.25	0.02	0.11	0.18	0.18	0.06	0.43	0.17	0.47	0.57	0.3	0	0.53	0.89
8503	0	0.05	0.01	0.01	0.01	0.01	0	0.01	0.01	0.02	0.01	0.02	0.01	0.01	0	0.01	0	0.02	0.01
8509	0.9	0.83	0.76	0.93	0.69	0.34	0.72	0.84	0.9	0.9	0.76	0.73	0.75	0.69	0.65	0.67	0.98	0.41	0.66
8512	0.95	0.93	0.98	0.97	0.93	0.64	0.94	0.93	1	0.67	0.99	0.78	0.89	0.88	0.98	0.85	0.85	0.97	0.99
8520	0.44	0.86	0.68	0.74	0.73	0.52	0.49	0.82	0.78	0.6	0.95	0.72	0.64	0.89	0.64	0.61	0.24	0.82	0.85
8521	0	0.02	0.04	0.02	0.07	0.01	0	0.01	0.05	0	0.05	0.08	0.05	0.06	0.03	0.07	0	0.1	0.16
8523	0.25	0.51	0.94	0.43	0.76	0.83	0.37	0.19	0.89	0.1	0.98	0.83	0.6	0.7		0.65	0.18	0.56	0.74
8534	0.94	0.96	0.99	0.79	0.8	0.73	0.91	0.87	0.94	0.84	0.92	0.69	0.52	0.89	0.83	0.61	0.95	0.7	0.66
8535	0.56	0.35	0.8	0.7	0.57	0.41	0.28	0.46	0.42	0.28	0.33	0.78	0.55	0.25	0.63	0.4	0.01	0.96	0.76
8536	0.91	0.98	0.99		1		0.92	0.96	1	0.99	1	0.99	0.98	1		1	0.91	0.99	0.98
8542	0.98	0.79	0.64	0.46	0.74	0.87	0.99	0.96	0.98	0.95	0.91	0.68	0.85	0.98	0.79	0.73	1	0.7	0.92
8556	0.78	0.38	0.53	0.53	0.35	0.89	0.72	0.69	0.64	0.92	0.51	0.8	0.59	0.23	0.81	0.67	0.87	0.75	0.85
8591	0.98	0.75	0.63	0.79	0.53	0.94	0.91	0.88	0.98	0.76	0.86	0.73	0.89	0.21	0.51	0.68	0.98	0.74	0.74
8596	0.11	0.67	0.33	0.43	0.3	0.33	0.12	0.32	0.26	0.37	0.13	0.44	0.25	0.28	0.32	0.24	0.31	0.36	0.5
8602	0.31	0.03	0.09	0.04	0.16	0.29	0.45	0.07	0	0.15	0.04	0.1	0.12	0.15	0.15	0.08	0.53	0.03	0.03
8603	0.99	0.96	0.89	0.97	0.74	0.5	0.99	0.93	1	0.96	0.89	0.73	0.86	0.77	0.77	0.74	1	0.35	0.93

Sentence Source	wiley.com	gov.au	abc.net.au	nationalgeographic.com	smh.com.au	oreilly.com	springer.com	edu	wikipedia.org	nejm.org	bbc.com	time.com	mit.edu	au.news.yahoo.com	skynews.com.au	economist.com	ncbi.nlm.nih.gov	nytimes.com	imdb.com
8608	0.16	0.3	1	0.97	0.94	0.66	0.16	0.57	0.79	0.39	0.96	0.9	0.73	0.82	0.97	0.78	0.04	0.94	1
8621	0.02	0.07	0.12	0.03	0.07		0.01	0.03	0.28		0.07	0.03	0.07	0.19	0.06	0.03	0	0.07	0.23
8623	0.41	0.41	0.45	0.4	0.66	0.89	0.44	0.56	0.61	0.29	0.57	0.82	0.58	0.62	0.73	0.86	0.27	0.75	0.87
8633	0.02	0.02	0.03	0.01	0.04	0.05	0.01	0.02	0.1	0.03	0.05	0.04	0.07	0.09	0.02	0.1	0.01	0.06	0.09
8641	0	0	0		0.01	0.01	0	0	0.06		0	0	0.01	0.11	0	0.03	0	0.02	0
8648	0.99	0.51	0.26	0.53	0.48	0.63	0.99	0.53	0.51	0.94	0.73	0.46	0.55	0.35	0.4	0.31	0.97	0.36	0.29
8655	0.06	0.05	0.04	0.08	0.11	0.03	0.07	0.03	0	0.1	0.02	0.09	0.08	0.05	0.01	0.1	0.16	0.06	0.02
8659	0.01	0.09	0.03	0.04	0.17	0.08	0.01	0.03	0.06	0.06	0.02	0.13	0.06	0.31	0.05	0.02	0	0.18	0.48
8681	0.14	0.28	0.16	0.6	0.25	0.28	0.12	0.25	0.52	0.24	0.33	0.18	0.5	0.01	0.5	0.13	0.16	0.24	0.98
8690	0.79	0.54	0.62	0.65	0.51	0.6	0.75	0.9	0.71	0.78	0.56	0.65	0.59	0.22	0.75	0.66	0.7	0.77	0.53
8705	0.01	0.05	0.04	0.15	0.06	0.03	0.01	0.01	0.02	0.1	0.14	0.09	0.04	0.01	0.02	0.04	0	0.19	0.19
8708	0.99	0.96	0.96	0.18	0.98	0.99	0.99	1	0.94	0.77	0.98	0.93	0.98	0.98	0.99	0.98	0.99	0.97	0.98
8720	0.73	0.76	0.12	0.54	0.2	0.17	0.75	0.74	0.47	0.76	0.15	0.05	0.66	0.04	0.13	0.24	0.77	0.19	0.07
8731	0.95	0.73	0.89	0.91	0.85	0.87	0.95	0.95	0.99	0.92	0.9	0.62	0.82	0.62	0.86	0.55	0.9	0.68	0.86
8736	0.05	0.6	0.18	0.92	0.81		0.06	0.47	0.35	0.13	0.68	0.95	0.39	0.51		0.95	0	0.77	0.9
8739	0.02	0.26	0.61	0.38	0.33	0.31	0.02	0.08	0.62	0.07	0.52	0.53	0.72	0.38	0.57	0.41	0	0.4	0.9
8740	0	0.25	0.01	0.02	0.06	0.18	0.01	0.03	0	0	0	0.11	0.1	0.17	0.09	0.03	0	0	0
8745	0.48	0.84	0.91	0.79	0.88	0.6	0.59	0.5	0.49	0.5	0.98	0.69	0.53	0.9	0.1	0.55	0.64	0.64	0.98
8748	0.04	0.28	0.18	0.07	0.38	0.11	0.04	0.06	0.15	0.5	0.04	0.4	0.09	0.43	0.45	0.18	0.06	0.48	0.57
8752	0.98	0.63	0.41	0.78	0.85	0.7	0.94	0.49	0.94	1	0.86	0.84	0.61	0.73		0.54	1	0.85	0.96
8754	0.91	0.97	0.95	0.95	0.94	0.51	0.92	0.91	0.97	0.59	0.96	0.87	0.8	0.82	0.95	0.73	0.92	0.87	0.73
8764	1	0.98	0.98	0.92	0.97	0.94	1	1	0.99	1	0.97	0.97	0.96	0.97	0.98	0.96	1	0.95	0.94
8774	0.42	0.42	0.64	0.58	0.52	0.37	0.39	0.37	0.5	0.17	0.78	0.5	0.37	0.42	0.65	0.46	0.23	0.53	0.53
8775	0.23	0.93	0.98	0.99	0.98	0.85	0.27	0.54	1	0.62	0.99	0.96	0.61	0.99	0.99	0.97	0.16	0.99	1
8781	0.03	0.1	0.09	0	0.17	0.02	0.04	0.01	0	0	0.02	0.06	0.07	0.16	0.01	0.14	0.05	0.08	0.02
8812	0.98	0.75	0.29	0.58	0.38	0.86	0.97	0.91	0.67	0.96	0.27	0.28	0.8	0.76	0.45	0.17	1	0.5	0.35
8829	0.54	0.45	0.75	0.32	0.85	0.63	0.46	0.56	0.51	0.51	0.91	0.69	0.68	0.68	0.8	0.68	0.77	0.65	0.81
8841	0.92	0.57	0.52	0.42	0.33	0.79	0.9	0.86	0	0.54	0.23	0.45	0.4	0.26	0.48	0.75	0.85	0.58	0.35
8851	0.75	0.97	0.92	0.89	0.83	0.97	0.63	0.92	0.99	0.92	0.91	0.89	0.88	0.79	0.85	0.73	0.93	0.94	0.94
8855	0.04	0.01	0.01	0.06	0.07	0.03	0.03	0.01	0	0.02	0	0.06	0.09	0.01	0.01	0.09	0	0.11	0.1
8857	0.09	0.42	0.12	0.18	0.26	0.12	0.17	0.1	0.48	0.01	0.32	0.31	0.27	0.39	0.67	0.38	0.05	0.28	0.21
8858	0.73	0.52	0.53	0.14	0.64	0.64	0.79	0.59	0.3	0.94	0.48	0.55	0.46	0.79	0.46	0.47	0.97	0.48	0.51
8859	0.03	0.04	0.09	0.18	0.21	0.35	0.06	0.12	0.08	0.03	0.18	0.18	0.41	0.06	0.24	0.11	0	0.16	0.42
8875	0.98	0.96	0.95	0.88	0.96	0.86	0.98	0.98	1	1	0.89	0.96	0.85	0.93	0.9	0.85	1	0.96	0.98
8879	0.45	0.3	0.23	0.33	0.24	0.82	0.48	0.44	0.44	0.13	0.19	0.25	0.61	0.15	0.14	0.33	0.17	0.22	0.36
8885	0.48	0.46	0.18	0.17	0.62		0.45	0.29	0.26		0.28	0.39	0.27	0.57	0.38	0.53	0.75	0.41	0.34
8889	0.05	0.06	0.03	0.1	0.03	0.06	0.03	0.04	0	0.25	0.04	0.06	0.14	0.03	0.02	0.04	0.02	0.05	0.02
8914	0.07	0.03	0.12	0.09	0.06	0.01	0.18	0.02	0.02	0.38	0.19	0.12	0.1	0.05	0.04	0.04	0.17	0.08	0.23
8937	0.7	0.55	0.1	0.39	0.03	0.6	0.58	0.93	0.43	0.76	0.2	0.2	0.56	0.2	0.27	0.13	0.86	0.21	0.34
8940	0.36	0.09	0.13	0.14	0.03	0.08	0.05	0.48	0.08	0.25	0.08	0.05	0.17	0.02	0.01	0.03	0.14	0.1	0
8949	0.57	0.68	0.21	0.09	0.36	0.56	0.85	0.52	0.6	0.83	0.14	0.44	0.43	0.13	0.18	0.39	1	0.51	0.09
8953	0.88	0.76	0.92	0.93	0.9	0.91	0.85	0.92	0.9	0.87	0.92	0.93	0.88	0.98	0.97	0.94	0.85	0.76	0.94
8954	0.56	0.69	0.65	0.8	0.58	0.69	0.47	0.75	0.99	0.67	0.8	0.77	0.64	0.49	0.83	0.6	0.68	0.81	0.8
8972	0.08	0.26	0.35	0.25	0.25	0.24	0.17	0.2	0.11	0.01	0.27	0.16	0.2	0.41	0.23	0.36	0	0.28	0.03
8974	0.09	0.22	0.33	0.63	0.23	0.69	0.07	0.2	0.58	0.09	0.18	0.39	0.44	0.2	0.13	0.36	0.03	0.23	0.92
8987	0.3	0.25	0.45	0.48	0.49	0.03	0.5	0.12	0.55	0.29	0.42	0.46	0.29	0.37	0.59	0.44	0.19	0.58	0.77
8989	0.93	0.94	0.74	0.31	0.69	0.87	0.87	0.93	0.97	0.51	0.59	0.68	0.81	0.63	0.55	0.58	0.99	0.84	0.48
8992	0	0.05	0.12	0.16	0.03	0.43	0	0.01	0.19	0	0.08	0.23	0.42	0.07	0.11	0.31	0	0.15	0.24
9005	0.08	0.41	0.47	0.26	0.41	0.35	0.05	0.25	1	0.22	0.51	0.58	0.35	0.65	0.66	0.46	0.03	0.95	0.99
9012	0.04	0.17	0.55	0.32	0.46	0.31	0.04	0.13	0.5	0.2	0.61	0.47	0.3	0.21	0.59	0.6	0.01	0.69	0.55
9014	0	0.01	0.02	0.05	0.01	0.01	0	0	0.03	0.05	0.01	0.03	0.01	0	0.01	0.02	0	0.02	0.02
9015	1	1	1	0.99	0.98	0.98	1	1	1	0.98	1	0.97	0.99	0.97	1	0.97	1	0.98	1
9018	0.19	0.58	0.88	0.87	0.7	0.5	0.25	0.52	0.5	0.25	0.55	0.55	0.43	0.73	0.41	0.85	0.04	0.63	0.63
9025	0.99	0.99	1	0.99	1		1	0.99	0.99	0.97	1	1	0.99	1		1	1	1	1
9055	1	1	1		0.99	0.99	1	1	1		1	1	0.99	0.97	1	0.98	1	1	1
9057	0.98	0.99	0.99	0.99	0.97	0.92	0.98	0.99	0.99	0.88	0.99	0.95	0.8	0.99	0.99	0.92	1	0.96	0.94
9062	0.41	0.66	0.55	0.19	0.46	0.25	0.49	0.59	0.43	0.41	0.39	0.45	0.49	0.71	0.65	0.36	0.44	0.52	0.5
9068	0.03	0.71	0.69	0.31	0.6	0.08	0.03	0.59	0.37	0.1	0.22	0.39	0.3	0.2	0.57	0.58	0.05	0.76	0.53
9073	0.9	0.46	0.23	0.24	0.23	0.23	0.93	0.56	0.24	0.6	0.24	0.31	0.65	0.46	0.3	0.42	0.92	0.37	0.12
9081	0.33	0.67	0.95		0.97		0.51	0.93	0.98		0.99	0.97	0.98	0.98	0.99	0.92	0.12	0.94	0.99
9083	0.11	0.07	0.04	0.16	0.06	0.04	0.13	0.04	0.36	0.02	0.04	0.09	0.23	0.14	0.1	0.08	0.1	0.06	0.29
9087	0.99	0.96	0.94	0.32	0.93	0.89	0.99	0.98	0.65	0.99	0.59	0.77	0.64	0.85	0.85	0.85	1	0.84	0.38
9110	0	0	0	0	0	0	0.01	0	0	0.06	0	0.01	0	0	0	0.03	0	0	0
9111	0.82	0.9	0.77	0.92	0.79	0.98	0.66	0.9	0.88	0.52	0.87	0.79	0.72	0.84	0.85	0.79	0.88	0.56	0.76
9119	0.93	0.98	1	0.99	0.99	0.96	0.98	0.99	1	0.95	1	0.99	0.95	0.98	0.99	0.99	0.99	1	1
9128	0.18	0.13	0.02	0.38	0.05	0.49	0.3	0.28	0.19	0.21	0.22	0.46	0.44	0.44		0.03	0.05	0.61	0.49
9131	0.97	0.64	0.54	0.42	0.55	0.76	0.94	0.79	0.14	0.69	0.42	0.47	0.73	0.86	0.58	0.61	0.99	0.51	0.17

Sentence Source	wiley.com	gov.au	abc.net.au	nationalgeographic.com	smh.com.au	oreilly.com	springer.com	edu	wikipedia.org	nejm.org	bbc.com	time.com	mit.edu	au.news.yahoo.com	skynews.com.au	economist.com	ncbi.nlm.nih.gov	nytimes.com	imdb.com
9136	0.19	0.01	0.09	0.05	0.06	0.08	0.28	0.04	0	0.16	0.06	0.12	0.08	0.24	0.1	0.2	0.12	0.04	0.01
9139	0.24	0.57	0.78	0.83	0.51	0.69	0.15	0.55	0.95	0.75	0.61	0.63	0.5	0.34	0.72	0.6	0.02	0.59	0.58
9146	0.12	0.19	0.14	0.11	0.03	0.16	0.12	0.18	0.51	0.06	0.03	0.05	0.16	0.02	0.03	0.1	0.04	0.07	0.12
9147	0.98	0.95	0.98	1	0.62		0.96	0.99	0.92	0.73	0.91	0.65	0.99	0.34		1	0.94	0.84	0.6
9148	0.95	0.74	0.29	0.81	0.33		0.91	0.82	0.92	0.97	0.44	0.6	0.33	0.22		0.39	0.94	0.67	0.27
9161	1	0.98	0.98	0.98	0.87	0.99	0.99	1	1	1	1	0.97	0.83	0.89	0.96	0.86	1	1	1
9163	0.01	0.04	0.18	0.04	0.03	0.25	0.01	0.03	0		0.02	0.1	0.11	0.02	0.08	0.05	0.03	0.03	0.03
9167	0.99	0.96	0.94	0.92	0.76	0.91	0.99	0.99	0.99	0.71	0.85	0.73	0.96	0.91	0.76	0.72	1	0.81	0.57
9168	1	1	1	1	0.99	1	1	1	1	0.99	1	0.98	1	0.98	0.99	0.92	1	1	1
9171	0.02	0	0.03	0.37	0.12	0.01	0.01	0.01	0	0.12	0.04	0.03	0.04	0	0.05	0.03	0.01	0.04	0
9172	0.88	0.76	0.97	0.95	0.97	0.77	0.86	0.95	0.98	0.96	0.98	0.97	0.64	0.94	0.93	0.95	0.97	0.96	0.86
9179	0.43	0.5	0.78	0.23	0.89	0.13	0.37	0.2	0.17	0.86	0.44	0.62	0.19	0.68		0.93	0.76	0.67	0.39
9181	0.4	0.21	0.12	0.02		0.63	0.3	0.22	0.43	0.22			0.35			0.01	0.13	0.9	0.53
9190	0.84	0.86	0.09	0.41	0.13	0.92	0.89	0.9	0.16	0.66	0.04	0.32	0.85	0.17	0.17	0.15	0.77	0.27	0.4
9213	0.3	0.03	0.03	0.15	0.06	0.05	0.25	0.02	0.05	0.58	0.02	0.07	0.1	0.17	0.02	0.16	0.18	0.05	0.02
9221	0.01	0	0.23	0.07	0.06	0.05	0.04	0.01	0.15	0	0.11	0.13	0.1	0.11	0.12	0.06	0	0.14	0.14
9224	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0
9226	0.94	0.94	0.91	0.81	0.95	0.93	0.96	0.97	0.99	0.96	0.89	0.97	0.92	0.9	0.97	0.95	0.98	0.95	0.91
9228	0.99	0.96	0.94	0.97	0.82	0.84	0.99	0.97	1	1	0.92	0.84	0.73	0.75	0.81	0.68	1	0.98	0.99
9234	0	0	0.04	0.06	0.02	0	0	0	0.01	0	0.01	0.02	0.02			0.07	0	0.01	0.02
9259	1	1	0.95	0.98	0.96	1	1	1	1	1	0.96	0.91	0.99	0.97	0.93	0.92	1	0.92	0.71
9280	0.12	0.81	0.78	0.74	0.64	0.62	0.2	0.49	0.93	0.57	0.86	0.48	0.51	0.75	0.84	0.55	0.04	0.69	0.85
9289	0.56	0.7	0.69	0.79	0.69	0.86	0.46	0.7	0.61	0.85	0.96	0.8	0.72	0.79	0.66	0.62	0.69	0.63	0.7
9309	0.81	0.94	0.93	0.82	0.86	0.9	0.79	0.91	0.91	0.54	0.82	0.77	0.73	0.69	0.95	0.97	0.8	0.87	0.79
9311	0.34	0.44	0.72	0.69	0.66	0.48	0.42	0.58	0.4	0.55	0.92	0.67	0.57	0.69	0.72	0.56	0.35	0.75	0.86
9319	0.02	0	0		0	0.17	0.02	0.03	0.06	0	0.01	0	0.11	0.02	0	0.01	0.01	0.01	0.01
9323	0	0.01	0	0	0	0	0	0	0	0	0.11	0.01	0.02	0.02	0.01	0.01	0	0.01	0
9328	0.58	0.31	0.26	0.44	0.36	0.2	0.53	0.69	0.24	0.6	0.3	0.4	0.45	0.26		0.58	0.89	0.33	0.1
9342	0.44	0.11	0.09	0.1	0.3	0.29	0.47	0.2	0.15	0.72	0.23	0.31	0.18	0.29	0.33	0.42	0.61	0.33	0.28
9352	0.01	0.01	0	0	0	0	0	0	0.01	0	0		0.1	0	0	0	0	0	0
9357	0.98	0.87	0.75	0.73	0.8	0.95	0.98	0.97	0.66	0.39	0.73	0.69	0.94	0.74	0.71	0.8	0.94	0.65	0.49
9366	0.08	0.23	0.76	0.43	0.51	0.57	0.09	0.2	0.06	0.13	0.56	0.63	0.36	0.3	0.69	0.66	0	0.52	0.06
9370	1	1	1	0.98	1	1	1	1	1	1	1	0.99	1	1	1	0.98	1	1	1
9372	0.42	0.24	0.28	0.28	0.17	0.14	0.38	0.27	0.3	0.42	0.41	0.33	0.54	0.42	0.19	0.25	0.93	0.35	0.28
9373	0.96	0.91	0.97	0.96	0.94	0.96	0.97	0.93	0.93	0.96	0.95	0.96	0.87	0.99	0.95	0.98	0.96	0.93	0.86
9378	0	0	0	0.04	0.01	0	0	0	0.01	0.01	0	0.01	0.01	0	0	0.01	0	0.01	0.07
9379	0.29	0.31	0.24	0.35	0.44	0.32	0.32	0.2	0.56	0.81	0.47	0.4	0.74	0.02	0.24	0.48	0.32	0.49	0.67
9394	0.98	0.9	0.32	0.73	0.74	0.97	0.97	0.97	0.7	0.98	0.46	0.83	0.79	0.54	0.76	0.83	0.94	0.65	0.56
9403	0.99	0.88	0.62	0.7	0.87	0.94	0.99	0.98	0.84	0.96	0.76	0.77	0.93	0.2	0.82	0.82	1	0.8	0.73
9415	0.98	0.66	0.93	0.92	0.83	0.65	0.96	0.9	0.99	0.89	0.95	0.88	0.72	0.96	0.81	0.76	1	0.97	0.98
9427	0.56	0.92	0.95	0.71	0.84	0.57	0.48	0.9	0.41	0.47	0.76	0.72	0.69	0.78	0.81	0.73	0.61	0.56	0.14
9434	0.2	0.35	0.29	0.73	0.17	0.13	0.27	0.23	0.2	0.06	0.12	0.16	0.16	0.11			0.15	0.26	0.11
9438	0.5	0.08	0.02	0.02	0.07	0.11	0.35	0.1	0.09	0.81	0.17	0.11	0.12	0.16		0.02	0.7	0.07	0.12
9452	0.69	0.29	0.03	0.14	0.08	0.19	0.64	0.37	0	0.62	0.07	0.08	0.3	0.27	0.05	0.13	0.75	0.07	0.16
9464	0.73	0.06	0.04	0.06	0.04	0.53	0.72	0.34	0.09	0.6	0.04	0.06	0.16	0.13		0.07	0.69	0.04	0.01
9467	0.17	0.38	0.24	0.09	0.17	0.27	0.29	0.28	0.03	0.15	0.11	0.5	0.23	0.52	0.25	0.13	0.19	0.14	0.08
9469	0.2	0.58	0.94	0.41	0.69	0.52	0.1	0.25	0.5	0.12	0.78	0.43	0.41	0.49	0.55	0.39	0.01	0.52	0.64
9482	0.92	0.94	0.9	0.89	0.92	0.92	0.98	0.98	1	0.94	1	0.94	0.88	0.97	0.92	0.91	0.99	0.96	0.99
9489	0.19	0.1	0.08		0.01		0.23	0.06	0.15	0.27	0.02	0.03	0.07	0.02			0.11		0.01
9490	0.97	0.69	0.78	0.71	0.77	0.6	0.97	0.77	0.59	0.85	0.86	0.68	0.64	0.9	0.85	0.41	1	0.7	0.64
9494	0.35	0.69	0.48	0.65	0.67	0.93	0.55	0.59	0.48	0.17	0.65	0.47	0.79	0.68	0.54	0.46	0.17	0.54	0.27
9508	0.72	0.95	0.99	0.98	0.96	0.81	0.59	0.94	1	0.7	1	0.96	0.76	0.9	0.98	0.91	0.63	1	0.99
9515	0.6	0.21	0.11	0.16	0.18	0.43	0.59	0.27	0.14	0.4	0.1	0.14	0.37	0.58	0.02	0.41	0.55	0.13	0.13
9519	0.12	0.26	0.12	0.3	0.1	0.07	0.16	0.04	0.15	0.28	0.12	0.07	0.25	0.02	0.25	0.28	0.01	0.1	0.03
9524	0.01	0.02					0.01	0.01	0.02	0.04			0.01	0			0	0	
9542	0.98	0.88	0.95	0.85	0.83	0.87	0.98	0.96	0.9	0.8	0.8	0.89	0.82	0.91	0.88	0.78	1	0.87	0.94
9556	0.65	0.7	0.61	0.64	0.44	0.38	0.54	0.53	0.81	0.51	0.9	0.54	0.49	0.44	0.71	0.42	0.62	0.65	0.18
9559	0.27	0.97	0.6	0.57	0.77	0.55	0.36	0.62	0.41	0.46	0.84	0.75	0.19	0.73	0.78	0.72	0.19	0.89	0.69
9573	0.65	0.83	0.59	0.27	0.33	0.69	0.72	0.89	0.98	0.59	0.25	0.62	0.7	0.16	0.61	0.66	0.58	0.67	0.64
9576	0.87	0.3	0.12	0.09	0.24	0.08	0.8	0.16	0	0.91	0.19	0.14	0.16	0.13	0.18	0.34	0.95	0.36	0
9582	0.03	0.15	0.13	0.14	0.15	0.41	0.07	0.05	0.16	0.01	0.17	0.21	0.52	0.06	0.5	0.35	0.01	0.42	0.18
9589	0.02	0.05	0.01	0.01		0.74	0.02	0.07	0.13	0.06	0.04		0.34	0.36		0.05	0.01	0.02	0.06
9597	0.58	0.3	0.09	0.37	0.28	0.36	0.61	0.48	0.47	0.25	0.28	0.34	0.3	0.2	0.29	0.34	0.66	0.43	0.19
9601	0.63	0.88	0.8	0.61	0.75	0.44	0.6	0.95	0.39	0.86	0.55	0.44	0.6	0.31	0.46	0.44	0.81	0.44	0.35
9618	0.01	0	0.01	0.01	0.02	0	0.01	0	0	0.09	0.01	0.01	0.03	0.01	0.01	0	0	0.03	0.08
9624	0	0.03	0.56	0.06	0.07	0.02	0	0	0.09	0.01	0.12	0.12	0.05	0.05	0.18	0.13	0	0.11	0.21

Sentence Source	wiley.com	gov.au	abc.net.au	nationalgeographic.com	smh.com.au	oreilly.com	springer.com	edu	wikipedia.org	nejm.org	bbc.com	time.com	mit.edu	au.news.yahoo.com	skynews.com.au	economist.com	ncbi.nlm.nih.gov	nytimes.com	imdb.com
9647	0.01	0.06	0.09	0.04	0.18	0.07	0.01	0.01	0	0.05	0.04	0.29	0.08	0.45	0.07	0.42	0	0.14	0.05
9648	0.96	0.31	0.31	0.63	0.38	0.91	0.96	0.44	0.63	0.92	0.85	0.6	0.72	0.79	0.44	0.5	0.94	0.18	0.63
9659	0.86	0.64	0.74	0.79	0.78	0.72	0.78	0.87	0.85	0.88	0.85	0.56	0.72	0.33	0.86	0.29	0.96	0.75	0.82
9673	0.54	0.66	0.32	0.44	0.45	0.51	0.51	0.48	0.44	0.48	0.21	0.42	0.36	0.21	0.35		0.72	0.58	0.09
9682	0.66	0.55	0.22		0.25	0.64	0.6	0.64	0.32	0.61	0.12	0.36	0.37	0.2	0.13	0.61	0.59	0.39	0.07
9684	0.91	0.45	0.38	0.69	0.2	0.11	0.51	0.42	0.31	0.46	0.01	0.22	0.26	0.27	0.03	0.25	0.96	0.23	0
9688	0.49	0.22	0.4	0.21	0.53	0.2	0.55	0.37	0.45	0.29	0.96	0.37	0.16	0.96	0.53	0.56	0.27	0.56	0.91
9689	0.06	0.05	0	0.12	0	0.75	0.21	0.21	0.22		0	0	0.84	0.02	0.01	0.01	0.05	0.01	0.01
9690	0.91	0.85	0.86	0.84	0.66	0.91	0.86	0.85	1	0.64	0.8	0.92	0.73	0.8	0.8	0.83	0.91	0.91	0.9
9695	0.03	0.04	0.02	0.13	0.05	0.08	0.04	0.03	0	0.15	0.05	0.08	0.13	0.17	0.04	0.15	0	0.08	0.01
9698	0.41	0.72	0.64	0.8	0.82	0.97	0.62	0.65	0.89	0.38	0.84	0.74	0.81	0.08	0.72	0.57	0.24	0.91	0.99
9699	0.41	0.25	0.02	0.33	0.01	0.04	0.33	0.05	0.15	0.04	0.04	0.04	0.29	0.05	0.03	0.03	0.15	0.08	0.12
9703	0.09	0.09	0.22	0.3	0.25		0.16	0.15	0.67	0.01	0.19	0.32	0.11	0.13	0.32	0.31	0.03	0.48	0.36
9708	0.07	0.08	0.1	0.16	0.17	0.25	0.16	0.07	0.06	0.1	0.28	0.16	0.22	0.13	0.16	0.19	0.07	0.12	0.05
9719	1	1	0.96	0.98	0.93	0.99	1	1	1	1	0.96	0.98	0.99	0.98	0.99	0.96	1	0.98	0.99
9720	0.69	0.78	0.53	0.7	0.57	0.92	0.77	0.85	0.95	0.67	0.66	0.76	0.76	0.61	0.57	0.76	0.81	0.78	0.85
9722	0.72	0.93	0.9	0.61	0.69	0.88	0.7	0.62	0.78	0.39	0.76	0.49	0.65	0.31	0.44	0.59	0.42	0.55	0.41
9725	0.32	0.04	0.01	0.1	0.01	0.1	0.23	0.01	0.01	0.13	0.01	0.01	0.21	0.05	0	0.05	0.17	0.04	0.02
9727	0.73	0.17	0.16	0.21	0.29	0.16	0.78	0.61	0.14	0.79	0.23	0.47	0.38	0.33	0.06	0.39	0.78	0.41	0.21
9731	0.96	0.86	0.86	0.97	0.78	0.81	0.97	0.96	0.97	1	0.77	0.79	0.3	0.5	0.79	0.82	0.98	0.79	0.63
9732	0.61	0.8	0.71	0.93	0.62	0.99	0.36	0.97	0.99	0.93	0.72	0.93	0.68	0.95	0.83	0.72	0.95	0.98	0.88
9739	0.17	0.82	0.72	0.88	0.85	0.72	0.22	0.69	1	0.28	0.97	0.92	0.78	0.78	0.85	0.74	0.19	0.93	0.97
9740	0.01	0.39	0.65	0.62	0.71	0.39	0.02	0.15	1	0.18	0.85	0.71	0.61	0.35	0.75	0.74	0	0.75	0.93
9758	0.8	0.96	0.83	0.94	0.73	0.93	0.72	0.92	1	0.48	0.88	0.82	0.96	0.38	0.76	0.85	0.56	0.92	0.97
9769	0.37	0.16	0.15	0.29	0.25	0.12	0.51	0.38	0.07	0.37	0.15	0.22	0.27	0.26	0.13	0.3	0.63	0.35	0.14
9780	0.04	0.02	0.09	0.01	0.03	0.02	0.01	0.01	0.04	0.16	0.03	0.04	0.02	0.13	0	0	0.08	0.01	0.01
9791	0.87	0.75	0.71	0.81	0.76		0.82	0.75	0.23	0.76	0.32	0.7	0.61	0.89	0.68	0.83	0.86	0.62	0.58
9797	0.63	0.81	0.94	0.79	0.87	0.83	0.86	0.72	1	0.53	0.99	0.93	0.89	0.91	0.87	0.96	0.17	0.96	0.98
9812	0.01	0.01	0.01	0.05	0.03	0.04	0.01	0.01	0	0	0.01	0.1	0.03	0.02	0.01	0.12	0	0.06	0.02
9819	0.72	0.81	0.47	0.5	0.47	0.97	0.65	0.9	1	0.61	0.26	0.83	0.86	0.76	0.62	0.84	0.54	0.83	0.95
9829	0.57	0.96	0.98	0.98	0.99	0.99	0.64	0.91	0.94	0.88	0.99	0.99	0.7	0.97	0.99	0.97	0.63	0.98	1
9831	0.01	0.03	0.19	0.05	0.04	0.18	0.03	0.08	0.12	0.06	0.02	0.05	0.22	0.04	0.02	0.1	0	0.09	0.16
9848	0.5	0.26	0.23	0.59	0.41	0.25	0.57	0.77	0.78	0.36	0.45	0.55	0.74	0.59	0.53	0.39	0.5	0.69	0.83
9861	0.69	0.84	0.83	0.65	0.77	0.42	0.76	0.55	0.35	0.61	0.67	0.53	0.62	0.9	0.65	0.8	0.9	0.6	0.21
9868	0.04	0	0.03	0.01	0.04	0.02	0.05	0	0	0.01	0.02	0.03	0.02	0.1	0.03	0.04	0.07	0.02	0
9876	0.8	0.96	0.32	0.62	0.39	0.9	0.92	0.84	0.71	0.42	0.56	0.53	0.81	0.72	0.45	0.62	0.73	0.6	0.77
9899	0.66	0.8	0.7	0.88	0.52	0.94	0.66	0.93	0.83	0.38	0.52	0.81	0.85	0.53	0.75	0.65	0.58	0.65	0.85
9901	0.82	0.82	0.59	0.53	0.65	0.89	0.6	0.86	1	0.4	0.54	0.72	0.75	0.43	0.47	0.74	0.23	0.9	1
9902	0.52	0.13	0.61	0.11	0.43	0.33	0.14	0.39	0.36	0.39	0.19	0.22	0.43	0.44	0.17	0.32	0.26	0.38	0.31
9919	0.09	0.85	0.64	0.85	0.84	0.65	0.07	0.33	0.82	0.14	0.77	0.77	0.53	0.63	0.8	0.73	0.04	0.85	0.92
9956	1	0.92	0.97	1	1	1	1	0.99	1	1	0.93	0.99	0.97	0.99	0.95	0.97	1	1	1
9965	0.05	0.08	0.34	0.03	0.16	0.06	0.16	0.33	0.02	0.13	0.14	0.05	0.32	0	0.07	0.18	0	0.15	0.36
9967	0.05	0.05	0.19	0.34	0.16	0.79	0.09	0.27	0.45	0.1	0.04	0.45	0.47	0.06	0.71	0.16	0.05	0.7	0.57
9974	0.19	0.27	0.17	0.36	0.24	0.43	0.18	0.09	0.51	0.05	0.23	0.24	0.37	0.39	0.1	0.28	0.11	0.16	0.3
9975	0.83	0.9	0.79	0.96	0.81	0.97	0.78	0.87	0.89	0.46	0.88	0.79	0.77	0.65	0.82	0.79	0.9	0.65	0.75
9979	0.99	0.98	0.64	0.98	0.34	0.97	0.99	0.99	1	0.94	0.62	0.56	0.89	0.65	0.48	0.37	1	0.6	0.78
9985	0.27	0.05	0.01	0	0.11	0.04	0.12	0.02	0.01	0.16	0.09	0.06	0.09	0.19	0.06	0.07	0.85	0.1	0.08
10003	0.03	0.08	0.04	0.03	0.09	0.04	0.03	0.02	0.03		0.01	0.1	0.01	0.03	0.04	0.27	0	0.08	0.03
10009	0.16	0.25	0.67	0.25	0.36	0.15	0.2	0.22	0.63	0.04	0.67	0.42	0.32	0.2	0.55	0.37	0.03	0.39	0.85
10012	0.02	0	0.35	0.16	0.09	0.02	0.01	0.01	0.11		0.13	0.11	0.04	0.18	0.1	0.14	0	0.1	0.07
10038	0.92	0.86	0.96	0.88	0.85	0.9	0.9	0.91	0.93	0.24	0.92	0.92	0.81	0.75	0.93	0.8	0.98	0.89	0.98
10040	0.41	0.96	0.92	0.8	0.86	0.94	0.49	0.88	1	0.38	0.97	0.92	0.7	0.94	0.79	0.84	0.81	0.91	0.97
10041	0.7	0.09	0.39	0.58	0.12	0.47	0.67	0.26	0.52	0.39	0.61	0.23	0.59	0.32	0.14	0.35	0.52	0.14	0.13
10047	0.94	0.88	0.64	0.95	0.72	0.92	0.95	0.52	0.75	0.67	0.69	0.33	0.5	0.7	0.74	0.45	0.99	0.39	0.26
10054	0.8	0.5	0.22	0.19	0.24	0.23	0.69	0.57	0.44	0.68	0.16	0.24	0.5	0.19	0.1	0.3	0.75	0.34	0.27
10059	0	0	0	0.01	0	0.03	0	0	0		0	0	0.03	0		0.01	0	0	0.01
10064	0.99	0.79	0.5	0.44	0.36	0.81	0.99	0.97	0.62	0.97	0.51	0.52	0.9	0.55	0.64	0.79	0.99	0.73	0.31
10086	0.27	0.04	0.09	0.09	0.07	0.01	0.3	0.14	0.14	0.72	0.03	0.14	0.1	0.02	0.04	0.1	0.42	0.17	0.1
10090	0.01	0	0.01		0.03		0.01	0.01	0.01	0.09	0.01	0.02	0.01	0.02		0.01	0.04	0.02	0.07
10102	0.05	0.46	0.61	0.12	0.24	0.19	0.04	0.06	0.29	0.17	0.6	0.51	0.23	0.67	0.12	0.51	0.01	0.67	0.76
10118	0.7	0.39	0.03	0.73	0.18	0.06	0.68	0.05	0	0.85	0.1	0.11	0.2	0.13	0.16	0.25	0.98	0.11	0.02
10134	0.55	0.3	0.34		0.13		0.69	0.37	0.15	0.35	0.08	0.17	0.78	0.1			0.63	0.17	0.03
10156	0.36	0.92	0.86	0.82	0.93	0.96	0.49	0.83	0.85	0.74	0.94	0.89	0.8	0.84	0.96	0.94	0.37	0.92	0.99
10177	0.24	0.12	0.23	0.58	0.15	0.52	0.32	0.13	0.32	0.29	0.33	0.21	0.48	0.22	0.19	0.36	0.35	0.23	0.68
10182	0.51	0.07	0.17	0.31	0.22	0.21	0.48	0.15	0	0.57	0.03	0.22	0.31	0.55	0.26	0.21	0.6	0.14	0.05
10209	0.17	0.01	0	0.03	0		0.12	0.03	0	0	0	0	0.1	0	0	0.01	0.06	0	0
10228	0.91	0.53	0.23	0.84	0.47	0.48	0.85	0.82	1	0.63	0.6	0.81	0.7	0.66	0.7	0.63	0.85	0.71	0.99

Sentence Source	**wiley.com**	**gov.au**	**abc.net.au**	**nationalgeographic.com**	**smh.com.au**	**oreilly.com**	**springer.com**	**edu**	**wikipedia.org**	**nejm.org**	**bbc.com**	**time.com**	**mit.edu**	**au.news.yahoo.com**	**skynews.com.au**	**economist.com**	**ncbi.nlm.nih.gov**	**nytimes.com**	**imdb.com**
10230	0.44	0.03	0.11	0.31	0.32	0.08	0.19	0.3	0.46	0.14	0.15	0.26	0.29	0.19	0.15	0.17	0.13	0.29	0.38
10246	0.03	0.14	0.18	0.35	0.34	0.16	0.01	0.13	0.36	0.09	0.64	0.45	0.15	0.28	0.34	0.32	0	0.25	1
10254	0.57	0.53	0.41	0.55	0.45	0.7	0.61	0.23	0.7	0.31	0.58	0.64	0.85	0.79	0.42	0.67	0.51	0.72	0.05
10257	0.19	0.09	0.05	0.38	0.05	0.01	0.1	0.05	0.34	0.12	0.06	0.12	0.24	0.04	0.07	0.06	0.11	0.13	0.01
10258	0.01	0.21	0.2	0.23	0.21	0.3	0.01	0.05	0.81	0.02	0.24	0.29	0.32	0.26	0.44	0.47	0	0.32	0.6
10267	0.07	0.07	0.27	0.1	0.04	0.19	0.06	0.01	0.01	0	0.03	0.11	0.11	0.03	0.06	0.04	0.02	0.04	0.04
10275	0.72	0.57	0.49	0.4	0.59	0.51	0.68	0.52	0.73	0.96	0.51	0.64	0.43	0.07	0.65	0.52	0.97	0.72	0.74
10279	0.97	0.7	0.9	0.99	0.84	0.99	0.95	0.94	0.79	0.87	0.98	0.94	0.88	0.99	0.98	0.9	0.99	0.91	0.98
10281	0.65	0.94	0.94	0.94	0.97	0.85	0.7	0.98	0.95	0.86	0.98	0.94	0.95	0.94	0.98	0.92	0.53	0.94	0.96
10285	0.29	0.08	0.23	0.84	0.36	0.58	0.28	0.3	0.16	0.76	0.6	0.47	0.4	0.91	0.37	0.34	0.44	0.39	0.57
10296	0.01	0.07	0.05	0.03	0.06	0.03	0.02	0.01	0.13	0.01	0.07	0.05	0.09	0.01	0.06	0.04	0	0.08	0.15
10335	0.01	0.03	0.12	0.28	0.08	0.1	0.02	0.08	0.1	0.03	0.12	0.15	0.32	0.09	0.14	0.15	0	0.19	0.27
10344	0.99	0.98	0.99	0.99	0.99	1	0.99	0.99	0.96	0.91	1	0.99	0.98	1	1	0.99	1	0.99	0.99
10350	0.67	0.91	0.78	0.75	0.67	0.58	0.76	0.94	0.96	0.89	0.65	0.72	0.61	0.75	0.76	0.66	0.87	0.72	0.25
10373	0.19	0.94	0.99	0.87	0.98	0.89	0.39	0.74	1	0.65	0.99	0.98	0.82	0.96	0.98	0.89	0.44	0.99	0.97
10379	0.04	0.01	0.01	0.01	0	0	0.05	0	0	0.03	0	0.02	0.03	0	0	0.02	0	0	0
10382	0.67	0.23	0.14	0.14	0.16	0.35	0.41	0.4	0.35	0.31	0.14	0.25	0.47	0.1	0.07	0.19	0.56	0.11	0.06
10383	0.45	0.35	0.3	0.4	0.64	0.88	0.52	0.71	0.44	0.17	0.3	0.48	0.77	0.7	0.16	0.86	0.12	0.65	0.17
10396	0.72	0.64	0.08	0.54	0.38	0.76	0.87	0.91	0.99	0.77	0.24	0.37	0.76	0.32	0.2	0.59	0.9	0.51	0.67
10398	0.18	0.23	0.13		0.27	0.44	0.28	0.14	0.65	0.01	0.13	0.08	0.34	0.35		0.07	0.03	0.16	0.61
10417	0.91	0.79	0.85	0.85	0.73	0.46	0.84	0.83	1	0.89	0.93	0.62	0.69	0.12	0.7	0.38	0.99	0.97	0.94
10419	0.5	0.07	0.03	0.13	0.15	0.03	0.4	0.32	0	0.23	0.02	0.13	0.11	0.06	0.04	0.26	0.21	0.02	0
10425	0.51	0.33	0.26	0.44	0.14	0.57	0.48	0.27	0.72	0.13	0.31	0.29	0.67	0.15	0.28	0.45	0.2	0.32	0.37
10426	0.14	0.08	0.01				0.24	0.11	0.02	0.02	0	0	0.1	0		0.02	0.15	0	0
10434	0	0.47	0.22	0.12	0.64	0.14	0	0.01	0.09		0.26	0.71	0.04	0.38	0.45	0.66	0	0.81	0.63
10466	0.98	0.78	0.51	0.45	0.35	0.84	0.99	0.97	0.64	0.96	0.48	0.49	0.9	0.55	0.64	0.68	0.99	0.76	0.45
10468	0.72	0.9	0.97	0.94	0.94	0.72	0.82	0.88	0.74	0.97	0.95	0.87	0.88	0.96	0.95	0.74	0.71	0.76	0.95
10481	0.17	0.22	0.64	0.27	0.42	0.75	0.16	0.29	0.87	0.2	0.62	0.53	0.33	0.22	0.49	0.43	0.01	0.28	0.48
10493	0.43	0.09	0.02	0.02	0.07	0.09	0.38	0.09	0.07	0.64	0.03	0.1	0.04	0.08	0.02	0.03	0.55	0.06	0.05
10500	0.89	0.37	0.6	0.44	0.28	0.47	0.69	0.77	0.66	0.43	0.44	0.49	0.42	0.44	0.34	0.28	0.96	0.54	0.5
10501	0	0.01	0.17	0.19	0.09	0.03	0	0.01	0.05	0	0.1	0.07	0.04	0.04	0.04	0.12	0	0.07	0.1
10504	0.74	0.03	0.07	0.13	0.07	0.07	0.36	0.02	0	0.05	0.06	0.07	0.09	0.01	0.02	0.25	0.56	0.06	0.1
10505	0.42	0.86	0.78	0.61	0.95	0.68	0.71	0.59	0.6	0.56	0.96	0.8	0.5	0.46	0.93	0.73	0.43	0.6	0.95
10523	0.85	0.99	0.97	0.99	0.99	0.99	0.86	0.97	0.99	0.94	0.97	1	0.81	0.99	0.99	0.98	0.99	0.99	1
10524	0.84	0.64	0.65	0.47	0.2	0.16	0.69	0.61	0.64	0.59	0.47	0.39	0.39	0.41	0.33	0.13	0.96	0.34	0.57
10528	0.54	0.81	0.97	0.85	0.89	0.69	0.77	0.88	1	0.78	0.99	0.83	0.85	0.83	0.95	0.96	0.14	0.97	0.99
10544	0.22	0.29	0.64	0.13	0.24	0.11	0.22	0.08	0.19	0.07	0.4	0.36	0.12	0.34	0.48	0.41	0.09	0.25	0.11
10545	0.96	0.73	0.49	0.56	0.91	0.53	0.86	0.88	0.64	0.73	0.79	0.46	0.78	0.61		0.77	0.98	0.62	0.68
10551	0.25	0.87	0.95	0.82	0.83	0.4	0.42	0.46	0.89	0.08	0.95	0.8	0.59	0.62	0.8	0.69	0.24	0.75	0.84
10552	0.03	0.06	0.03	0.01	0.04	0.01	0.03	0.02	0.01	0.54	0.01	0.04	0.02	0.04	0.01	0.02	0.04	0.04	0.02
10577	0.6	0.95	0.24	0.83	0.93	0.79	0.65	0.94	0.75	0.57	0.75	0.9	0.87	0.91	0.95	0.95	0.72	0.92	0.85
10578	0.1	0.07	0.2	0.09	0.24	0.4	0.03	0.1	0.13	0.02	0.31	0.13	0.29	0.35	0.28	0.28	0.01	0.4	0.42
10584	0.96	0.99	0.98	0.93	0.96	0.85	0.9	1	0.99	0.99	0.98	0.99	0.76	0.98	0.99	0.96	0.94	0.97	0.95
10597	0.89	0.69	0.09	0.27	0.11	0.2	0.88	0.23	0.35	0.82	0.02	0.09	0.41	0.04	0.1	0.05	0.98	0.17	0.12
10599	0.65	0.46	0.68	0.77	0.51	0.92	0.76	0.75	0.76	0.35	0.59	0.77	0.81	0.6	0.6	0.8	0.59	0.74	0.96
10605	0.7	0.47	0.29	0.85	0.61	0.16	0.81	0.42	0.3	0.85	0.58	0.42	0.33	0.46	0.42	0.24	0.96	0.54	0.48
10606	0.22	0.79	0.77	0.76	0.72	0.62	0.2	0.59	0.96	0.75	0.66	0.73	0.82	0.38	0.71	0.6	0.63	0.81	0.84
10614	0.29	0.27	0.09		0.02	0.15	0.05	0.15	0.16	0.17	0.02	0.07	0.28	0.04	0.02	0.04	0.04	0.03	0.02
10622	0.99	0.92	0.88	0.92	0.81	0.82	0.99	0.91	1	0.99	0.94	0.77	0.41	0.97	0.71	0.65	1	0.98	0.97
10627	0.06	0.33	0.2	0.4	0.18	0.36	0.08	0.22	0.72	0.08	0.44	0.19	0.24	0.12	0.04	0.12	0.08	0.15	0.27
10629	0.77	0.65	0.38	0.41	0.08	0.15	0.86	0.85	0.59	0.85	0.22	0.43	0.64	0.67	0.13	0.27	0.98	0.4	0.08
10637	1	0.99	0.98	0.98	0.97	0.99	1	1	1	0.99	0.98	0.92	0.97	0.97	0.99	0.96	1	0.9	0.98
10639	0.46	0.09	0.06	0.08	0.05	0.07	0.2	0.14	0	0.32	0.01	0.04	0.13	0.25	0.01	0.03	0.58	0.08	0
10641	0.98	0.6	0.36	0.86	0.52	0.17	0.97	0.89	0.84	0.82	0.81	0.65	0.63	0.42	0.75	0.52	0.97	0.6	0.63
10642	0.54	0.66	0.81	0.96	0.75		0.51	0.7	0.67	0.39	0.64	0.92	0.64	0.28	0.68	0.94	0.48	0.84	0.94
10643	0.07	0.01	0	0.07	0.03	0.09	0.09	0.03	0.05	0.16	0.01	0.14	0.2	0.07	0.02	0.12	0.23	0.15	0.06
10645	0.97	0.44	0.13	0.3	0.03	0.78	0.98	0.86	0.74	0.85	0.2	0.17	0.8	0		0.05	0.99	0.38	0.92
10646	0.65	0.65	0.85	0.81	0.99	0.94	0.64	0.9	1	0.9	0.94	0.89	0.73	0.9	0.92	0.84	0.69	0.85	0.86
10650	0.19	0.55	0.71	0.52	0.93	0.53	0.29	0.48	0.49	0.28	0.61	0.58	0.46	0.75	0.68	0.61	0.37	0.99	0.98
10692	0	0	0	0.02	0	0	0	0	0	0.07	0	0.01	0	0	0.01	0.01	0.01	0.01	0.02
10703	0.96	0.82	0.96	0.93	0.93	0.65	0.65	0.9	1	0.67	0.93	0.9	0.78	0.79	0.96	0.94	0.28	0.97	0.98
10704	0.01	0.01	0	0.01	0.01	0.02	0.01	0.01	0	0.02	0.01	0.01	0.03	0	0.01	0.01	0	0.02	0.02

B.1.2 $P(e2 \rightarrow e1)$

Sentence Source	wiley.com	gov.au	abc.net.au	nationalgeographic.com	smh.com.au	oreilly.com	springer.com	edu	wikipedia.org	nejm.org	bbc.com	time.com	mit.edu	au.news.yahoo.com	skynews.com.au	economist.com	ncbi.nlm.nih.gov	nytimes.com	imdb.com
8005	0.03	0.30	0.10	0.02	0.16	0.01	0.05	0.06	0.18	0.12	0.02	0.07	0.10	0.01	0.02	0.07	0.02	0.09	0.03
8027	0.47	0.32	0.13	0.20	0.16	0.24	0.60	0.35	0.07	0.51	0.05	0.16	0.34	0.59	0.17	0.43	0.65	0.05	0.24
8031	0.26	0.29	0.45	0.71	0.31	0.87	0.40	0.41	0.62	0.41	0.71	0.74	0.58	0.89	0.60	0.63	0.34	0.39	0.69
8041	0.30	0.78	0.71	0.78	0.76	0.72	0.35	0.76	1.00	0.54	0.86	0.93	0.68	0.78	0.79	0.68	0.52	0.80	0.95
8043	0.01	0.04	0.04	0.13	0.06	0.01	0.01	0.01	0.00	0.01	0.06	0.19	0.05	0.17	0.14	0.23	0.00	0.09	0.06
8055	0.36	0.34	0.82	0.83	0.57	0.40	0.37	0.41	0.99	0.41	0.75	0.61	0.62	0.47	0.77	0.45	0.41	0.77	0.97
8058	0.10	0.07	0.06	0.07	0.03	0.16	0.11	0.04	0.00	0.05	0.05	0.06	0.12	0.03	0.01	0.04	0.01	0.05	0.05
8073	0.97	0.26	0.25	0.48	0.37	0.82	0.92	0.79	0.06	0.96	0.10	0.55	0.53	0.51	0.17	0.55	0.99	0.52	0.00
8083	0.09	0.17	0.24	0.15	0.13	0.13	0.06	0.13	0.08	0.12	0.16	0.13	0.16	0.13	0.10	0.11	0.04	0.13	0.14
8105	0.32	0.83	0.21	0.48	0.10	0.50	0.59	0.60	0.22	0.50	0.50	0.48	0.41	0.58	0.37	0.51	0.67	0.39	0.47
8107	0.73	0.99	0.91	0.87	0.97	0.92	0.84	0.96	1.00	0.92	0.97	0.95	0.76	0.98	0.99	0.87	0.77	0.98	0.96
8108	0.99	0.99	0.98	0.98	0.97	0.99	1.00	1.00	1.00	1.00	0.98	0.98	0.95	0.86	0.99	0.95	1.00	0.94	0.78
8116	0.67	0.66	0.42	0.38	0.39	0.86	0.64	0.65	0.64	0.72	0.57	0.47	0.38	0.50	0.64	0.55	0.80	0.75	0.53
8118	0.40	0.59	0.27	0.16	0.12	0.52	0.39	0.30	0.39	0.15	0.15	0.16	0.42	0.30	0.46	0.09	0.56	0.39	0.09
8159	0.78	0.17	0.73	0.66	0.16	0.39	0.71	0.67	0.39	0.15	0.29	0.27	0.72	0.25	0.28	0.39	0.95	0.23	0.08
8175	0.33	0.16	0.04	0.51	0.13	0.13	0.47	0.10	0.00	0.15	0.08	0.34	0.19	0.28	0.16	0.23	0.66	0.14	0.00
8191	0.97	0.72	0.96	0.93	0.82	0.68	0.95	0.74	1.00	0.82	0.93	0.57	0.83	0.46	0.70	0.79	0.83	0.85	0.93
8204	0.99	1.00	0.99	0.94	0.99	1.00	1.00	1.00	1.00	0.99	1.00	0.99	0.99	0.99	1.00	1.00	0.99	0.99	1.00
8219	0.03	0.42	0.36	0.27	0.41	0.28	0.09	0.09	0.00	0.03	0.06	0.39	0.16	0.30	0.27	0.29	0.01	0.09	0.39
8234	0.99	0.93	0.97	0.97	0.91	0.81	0.99	0.91	0.96	0.86	0.96	0.93	0.89	0.97	0.96	0.91	1.00	0.92	0.92
8236	0.13	0.51	0.30	0.32	0.39	0.00	0.07	0.04	0.05	0.19	0.15	0.18	0.02	0.12	0.11	0.10	0.14	0.22	0.22
8239	0.90	0.88	0.60	0.99	0.81	0.97	0.93	0.68	0.90	0.52	0.76	0.54	0.76	0.82	0.79	0.69	0.93	0.57	0.69
8240	0.02	0.12	0.39	0.06	0.35	0.07	0.05	0.03	0.01	0.31	0.45	0.11	0.13	0.04	0.40	0.08	0.00	0.43	0.60
8253	0.01	0.13	0.13	0.03	0.08	0.00	0.02	0.01	0.11	0.00	0.20	0.20	0.03	0.28	0.35	0.20	0.00	0.25	0.03
8257	0.99	0.84	0.96	0.91	0.97	0.97	0.99	0.99	0.99	1.00	0.98	0.99	0.92	0.92	0.99	0.98	1.00	0.98	1.00
8265	1.00	1.00	0.96	0.84	0.97	0.92	1.00	0.99	0.95	0.99	0.92	0.89	0.94	0.89	0.91	0.97	1.00	0.70	0.78
8288	0.01	0.12	0.60	0.31	0.71	0.08	0.01	0.10	0.50	0.26	0.80	0.48	0.20	0.18	0.80	0.31	0.00	0.39	0.96
8312	0.80	0.89	0.84	0.97	0.98	0.96	0.76	0.98	1.00	0.93	0.97	0.93	0.86	0.95	0.99	0.76	0.89	0.96	1.00
8334	0.96	0.32	0.38	0.95	0.81	0.58	0.96	0.85	0.91	0.95	0.68	0.41	0.72	0.33	0.94	0.28	0.99	0.51	0.79
8357	0.76	0.35	0.09	0.37	0.71	0.36	0.80	0.65	0.25	0.40	0.23	0.11	0.50	0.20	0.25	0.19	0.75	0.22	0.18
8361	0.48	0.82	0.53	0.69	0.82	0.83	0.48	0.73	0.72	0.89	0.80	0.75	0.48	0.27	0.76	0.65	0.63	0.79	0.91
8373	0.99	0.98	0.96	0.84	0.83	0.53	0.99	0.94	0.88	0.93	0.83	0.73	0.81	0.98	0.87	0.82	1.00	0.75	0.40
8377	0.09	0.10	0.00		0.00		0.10	0.03	0.09	0.09	0.01	0.02	0.01	0.02	0.00	0.01	0.11	0.02	0.00
8382	0.98	0.98	0.90	0.75	0.92	0.87	0.98	0.92	0.91	0.97	0.91	0.78	0.84	0.98	0.86	0.81	1.00	0.85	0.74
8402	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.03	0.51	0.02	0.02	0.00	0.01	0.02	0.00	0.01	0.00
8403	0.99	0.94	0.69	0.77	0.56	0.85	0.99	0.92	0.71	0.95	0.70	0.50	0.32	0.74	0.51	0.58	1.00	0.54	0.23
8404	1.00	0.97	0.98	0.95	0.97	0.91	1.00	1.00	1.00	1.00	0.99	0.96	0.94	0.90	0.98	0.90	1.00	1.00	1.00
8405	0.04	0.09	0.00	0.02	0.01	0.00	0.07	0.04	0.00	0.16	0.00	0.03	0.04	0.14	0.01	0.11	0.02	0.01	0.00
8409	0.03	0.01	0.01	0.01	0.01	0.03	0.02	0.00	0.00	0.04	0.00	0.00	0.07	0.01	0.00	0.02	0.01	0.00	0.00
8417	0.00	0.00	0.00		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00			0.00	0.00	
8439	0.94	0.90	0.79	0.90	0.63	0.94	0.94	0.98	0.68	0.90	0.73	0.94	0.85	0.98	0.75	0.93	0.97	0.89	0.84
8455	0.99	0.95	0.92	0.96	0.96	0.90	1.00	0.85	0.51	0.99	0.63	0.89	0.92	0.87	0.90	0.89	1.00	0.85	0.95
8471	0.64	0.13	0.28	0.02	0.22	0.23	0.28	0.59	0.49	0.41	0.09	0.16	0.48	0.53	0.31	0.38	0.51	0.33	0.08
8473	0.19	0.12	0.52	0.57	0.24	0.39	0.16	0.22	0.44	0.14	0.58	0.50	0.45	0.71	0.51	0.58	0.04	0.32	0.69
8476	0.01	0.56	0.50	0.09	0.43	0.49	0.01	0.08	0.30	0.03	0.36	0.59	0.24	0.76	0.51	0.65	0.00	0.57	0.48
8486	0.01	0.24	0.13	0.09	0.18	0.31	0.02	0.14	0.00	0.03	0.07	0.21	0.26	0.14	0.25	0.48	0.00	0.04	0.03
8489	0.06	0.07	0.37	0.26	0.36	0.09	0.09	0.04	0.00	0.23	0.05	0.41	0.16	0.92	0.36	0.46	0.03	0.15	0.17
8493	0.95	0.99	0.81	0.89	0.97	0.84	0.84	0.96	0.96	0.98	0.97	0.95	0.76	0.99	0.97	0.94	0.94	0.96	0.98
8494	0.01	0.13	0.50	0.26	0.91	0.07	0.02	0.04	0.42	0.21	0.26	0.46	0.08	0.80	0.57	0.45	0.00	0.64	0.73
8498	0.67	0.60	0.65	0.78	0.51	0.42	0.57	0.61	0.72	0.55	0.47	0.47	0.67	0.37	0.38	0.47	0.61	0.48	0.54
8500	0.98	0.55	0.82	0.78	0.53	0.75	0.98	0.89	0.82	0.82	0.94	0.57	0.83	0.53	0.43	0.70	1.00	0.47	0.11
8503	1.00	0.95	0.99	0.99	0.99	0.99	1.00	0.99	0.99	0.98	0.99	0.98	0.99	0.99	1.00	0.99	1.00	0.98	0.99
8509	0.10	0.17	0.24	0.07	0.31	0.66	0.28	0.16	0.10	0.10	0.24	0.27	0.25	0.31	0.35	0.33	0.02	0.59	0.34
8512	0.05	0.07	0.02	0.03	0.07	0.36	0.06	0.07	0.00	0.33	0.01	0.22	0.11	0.12	0.02	0.15	0.15	0.03	0.01
8520	0.56	0.14	0.32	0.26	0.27	0.48	0.51	0.18	0.22	0.40	0.05	0.28	0.36	0.11	0.36	0.39	0.76	0.18	0.15
8521	1.00	0.98	0.96	0.98	0.93	0.99	1.00	0.99	0.95	1.00	0.95	0.92	0.95	0.94	0.97	0.93	1.00	0.90	0.84
8523	0.75	0.49	0.06	0.57	0.24	0.17	0.63	0.81	0.11	0.90	0.02	0.17	0.40	0.30		0.35	0.82	0.44	0.26
8534	0.06	0.04	0.01	0.21	0.20	0.27	0.09	0.13	0.06	0.16	0.08	0.31	0.48	0.11	0.17	0.39	0.05	0.30	0.34
8535	0.44	0.65	0.20	0.30	0.43	0.59	0.72	0.54	0.58	0.72	0.67	0.22	0.45	0.75	0.37	0.60	0.99	0.04	0.24
8536	0.09	0.02	0.01		0.00		0.08	0.04	0.00	0.01	0.00	0.01	0.02	0.00		0.00	0.09	0.01	0.02
8542	0.02	0.21	0.36	0.54	0.26	0.13	0.01	0.04	0.02	0.05	0.09	0.32	0.15	0.02	0.21	0.27	0.00	0.30	0.08
8556	0.22	0.62	0.47	0.47	0.65	0.11	0.28	0.31	0.36	0.08	0.49	0.20	0.41	0.77	0.19	0.33	0.13	0.25	0.15
8591	0.02	0.25	0.37	0.21	0.47	0.06	0.09	0.12	0.02	0.24	0.14	0.27	0.11	0.79	0.49	0.32	0.02	0.26	0.26
8596	0.89	0.33	0.67	0.57	0.70	0.67	0.88	0.68	0.74	0.63	0.87	0.56	0.75	0.72	0.68	0.76	0.69	0.64	0.50
8602	0.69	0.97	0.91	0.96	0.84	0.71	0.55	0.93	1.00	0.85	0.96	0.90	0.88	0.85	0.85	0.92	0.47	0.97	0.97
8603	0.01	0.04	0.11	0.03	0.26	0.50	0.01	0.07	0.00	0.04	0.11	0.27	0.14	0.23	0.23	0.26	0.00	0.65	0.07

Sentence Source	wiley.com	gov.au	abc.net.au	nationalgeographic.com	smh.com.au	oreilly.com	springer.com	edu	wikipedia.org	nejm.org	bbc.com	time.com	mit.edu	au.news.yahoo.com	skynews.com.au	economist.com	ncbi.nlm.nih.gov	nytimes.com	imdb.com
8608	0.84	0.70	0.00	0.03	0.06	0.34	0.84	0.43	0.21	0.61	0.04	0.10	0.27	0.18	0.03	0.22	0.96	0.06	0.00
8621	0.98	0.93	0.88	0.97	0.93		0.99	0.97	0.72		0.93	0.97	0.93	0.81	0.94	0.97	1.00	0.93	0.77
8623	0.59	0.59	0.55	0.60	0.34	0.11	0.56	0.44	0.39	0.71	0.43	0.18	0.42	0.38	0.27	0.14	0.73	0.25	0.13
8633	0.98	0.98	0.97	0.99	0.96	0.95	0.99	0.98	0.90	0.97	0.95	0.96	0.93	0.91	0.98	0.90	0.99	0.94	0.91
8641	1.00	1.00	1.00		0.99	0.99	1.00	1.00	0.94		1.00	1.00	0.99	0.89	1.00	0.97	1.00	0.98	1.00
8648	0.01	0.49	0.74	0.47	0.52	0.37	0.01	0.47	0.49	0.06	0.27	0.54	0.45	0.65	0.60	0.69	0.03	0.64	0.71
8655	0.94	0.95	0.96	0.92	0.89	0.97	0.93	0.97	1.00	0.90	0.98	0.91	0.92	0.95	0.99	0.90	0.84	0.94	0.98
8659	0.99	0.91	0.97	0.96	0.83	0.92	0.99	0.97	0.94	0.94	0.98	0.87	0.94	0.69	0.95	0.98	1.00	0.82	0.52
8681	0.86	0.72	0.84	0.40	0.75	0.72	0.88	0.75	0.48	0.76	0.67	0.82	0.50	0.99	0.50	0.87	0.84	0.76	0.02
8690	0.21	0.46	0.38	0.35	0.49	0.40	0.25	0.10	0.29	0.22	0.44	0.35	0.41	0.78	0.25	0.34	0.30	0.23	0.47
8705	0.99	0.95	0.96	0.85	0.94	0.97	0.99	0.99	0.98	0.90	0.86	0.91	0.96	0.99	0.98	0.96	1.00	0.81	0.81
8708	0.01	0.04	0.04	0.82	0.02	0.01	0.01	0.00	0.06	0.23	0.02	0.07	0.02	0.02	0.01	0.02	0.01	0.03	0.02
8720	0.27	0.24	0.88	0.46	0.80	0.83	0.25	0.26	0.53	0.24	0.85	0.95	0.34	0.96	0.87	0.76	0.23	0.81	0.93
8731	0.05	0.27	0.11	0.09	0.15	0.13	0.05	0.05	0.01	0.08	0.10	0.38	0.18	0.38	0.14	0.45	0.10	0.32	0.14
8736	0.95	0.40	0.82	0.08	0.19		0.94	0.53	0.65	0.87	0.32	0.05	0.61	0.49		0.05	1.00	0.23	0.10
8739	0.98	0.74	0.39	0.62	0.67	0.69	0.98	0.92	0.38	0.93	0.48	0.47	0.28	0.62	0.43	0.59	1.00	0.60	0.10
8740	1.00	0.75	0.99	0.98	0.94	0.82	0.99	0.97	1.00	1.00	1.00	0.89	0.90	0.83	0.91	0.97	1.00	1.00	1.00
8745	0.52	0.16	0.09	0.21	0.12	0.40	0.41	0.50	0.51	0.50	0.02	0.31	0.47	0.10	0.90	0.45	0.36	0.36	0.02
8748	0.96	0.72	0.82	0.93	0.62	0.89	0.96	0.94	0.85	0.50	0.96	0.60	0.91	0.57	0.55	0.82	0.94	0.52	0.43
8752	0.02	0.37	0.59	0.22	0.15	0.30	0.06	0.51	0.06	0.00	0.14	0.16	0.39	0.27		0.46	0.00	0.15	0.04
8754	0.09	0.03	0.05	0.05	0.06	0.49	0.08	0.09	0.03	0.41	0.04	0.13	0.20	0.18	0.05	0.27	0.08	0.13	0.27
8764	0.00	0.02	0.02	0.08	0.03	0.06	0.00	0.00	0.01	0.00	0.03	0.03	0.04	0.03	0.02	0.04	0.00	0.05	0.06
8774	0.58	0.58	0.36	0.42	0.48	0.63	0.61	0.63	0.50	0.83	0.22	0.50	0.63	0.58	0.35	0.54	0.77	0.47	0.47
8775	0.77	0.07	0.02	0.01	0.02	0.15	0.73	0.46	0.00	0.38	0.01	0.04	0.39	0.01	0.01	0.03	0.84	0.01	0.00
8781	0.97	0.90	0.91	1.00	0.83	0.98	0.96	0.99	1.00	1.00	0.98	0.94	0.93	0.84	0.99	0.86	0.95	0.92	0.98
8812	0.02	0.25	0.71	0.42	0.62	0.14	0.03	0.09	0.33	0.04	0.73	0.72	0.20	0.24	0.55	0.83	0.00	0.50	0.65
8829	0.46	0.55	0.25	0.68	0.15	0.37	0.54	0.44	0.49	0.49	0.09	0.31	0.32	0.32	0.20	0.32	0.23	0.35	0.19
8841	0.08	0.43	0.48	0.58	0.67	0.21	0.10	0.14	1.00	0.46	0.77	0.55	0.60	0.74	0.52	0.25	0.15	0.42	0.65
8851	0.25	0.03	0.08	0.11	0.17	0.03	0.37	0.08	0.01	0.08	0.09	0.11	0.12	0.21	0.15	0.27	0.07	0.06	0.06
8855	0.96	0.99	0.99	0.94	0.93	0.97	0.97	0.99	1.00	0.98	1.00	0.94	0.91	0.99	0.99	0.91	1.00	0.89	0.90
8857	0.91	0.58	0.88	0.82	0.74	0.88	0.83	0.90	0.52	0.99	0.68	0.69	0.73	0.61	0.33	0.62	0.95	0.72	0.79
8858	0.27	0.48	0.47	0.86	0.36	0.36	0.21	0.41	0.70	0.06	0.52	0.45	0.54	0.21	0.54	0.53	0.03	0.52	0.49
8859	0.97	0.96	0.91	0.82	0.79	0.65	0.94	0.88	0.92	0.97	0.82	0.82	0.59	0.94	0.76	0.89	1.00	0.84	0.58
8875	0.02	0.04	0.05	0.12	0.04	0.14	0.02	0.02	0.00	0.00	0.11	0.04	0.15	0.07	0.10	0.15	0.00	0.04	0.02
8879	0.55	0.70	0.77	0.67	0.76	0.18	0.52	0.56	0.56	0.87	0.81	0.75	0.39	0.85	0.86	0.67	0.83	0.78	0.64
8885	0.52	0.54	0.82	0.83	0.38		0.55	0.71	0.74		0.72	0.61	0.73	0.43	0.62	0.47	0.25	0.59	0.66
8889	0.95	0.94	0.97	0.90	0.97	0.94	0.97	0.96	1.00	0.75	0.96	0.94	0.86	0.97	0.98	0.96	0.98	0.95	0.98
8914	0.93	0.97	0.88	0.91	0.94	0.99	0.82	0.98	0.98	0.62	0.81	0.88	0.90	0.95	0.96	0.96	0.83	0.92	0.77
8937	0.30	0.45	0.90	0.61	0.97	0.40	0.42	0.07	0.57	0.24	0.80	0.80	0.44	0.80	0.73	0.87	0.14	0.79	0.66
8940	0.64	0.91	0.87	0.86	0.97	0.92	0.95	0.52	0.92	0.75	0.92	0.95	0.83	0.98	0.99	0.97	0.86	0.90	1.00
8949	0.43	0.32	0.79	0.91	0.64	0.44	0.15	0.48	0.40	0.17	0.86	0.56	0.57	0.87	0.82	0.61	0.00	0.49	0.91
8953	0.12	0.24	0.08	0.07	0.10	0.09	0.15	0.08	0.10	0.13	0.08	0.07	0.12	0.02	0.03	0.06	0.15	0.24	0.06
8954	0.44	0.31	0.35	0.20	0.43	0.31	0.53	0.25	0.01	0.33	0.20	0.23	0.36	0.51	0.17	0.40	0.32	0.19	0.20
8972	0.92	0.74	0.65	0.75	0.75	0.76	0.83	0.80	0.89	0.99	0.73	0.84	0.80	0.59	0.77	0.64	1.00	0.72	0.97
8974	0.91	0.78	0.67	0.37	0.77	0.31	0.93	0.80	0.42	0.91	0.82	0.61	0.56	0.80	0.87	0.64	0.97	0.77	0.08
8987	0.70	0.75	0.55	0.52	0.51	0.97	0.50	0.88	0.45	0.71	0.58	0.54	0.71	0.63	0.41	0.56	0.81	0.42	0.23
8989	0.07	0.06	0.26	0.69	0.31	0.13	0.13	0.07	0.03	0.49	0.41	0.32	0.19	0.37	0.45	0.42	0.01	0.16	0.52
8992	1.00	0.95	0.88	0.84	0.97	0.57	1.00	0.99	0.81	1.00	0.92	0.77	0.58	0.93	0.89	0.69	1.00	0.85	0.76
9005	0.92	0.59	0.53	0.74	0.59	0.65	0.95	0.75	0.00	0.78	0.49	0.42	0.65	0.35	0.34	0.54	0.97	0.05	0.01
9012	0.96	0.83	0.45	0.68	0.54	0.69	0.96	0.87	0.50	0.80	0.39	0.53	0.70	0.79	0.41	0.40	0.99	0.31	0.45
9014	1.00	0.99	0.98	0.95	0.99	0.99	1.00	1.00	0.97	0.95	0.99	0.97	0.99	1.00	0.99	0.98	1.00	0.98	0.98
9015	0.00	0.00	0.00	0.01	0.02	0.02	0.00	0.00	0.00	0.02	0.00	0.03	0.01	0.03	0.00	0.03	0.00	0.02	0.00
9018	0.81	0.42	0.12	0.13	0.30	0.50	0.75	0.48	0.50	0.75	0.45	0.45	0.57	0.27	0.59	0.15	0.96	0.37	0.37
9025	0.01	0.01	0.00	0.01	0.00		0.00	0.01	0.01	0.03	0.00	0.00	0.01	0.00		0.00	0.00	0.00	0.00
9055	0.00	0.00	0.00		0.01	0.01	0.00	0.00	0.00		0.00	0.00	0.01	0.03	0.00	0.02	0.00	0.00	0.00
9057	0.02	0.01	0.01	0.01	0.03	0.08	0.02	0.01	0.01	0.12	0.01	0.05	0.20	0.01	0.01	0.08	0.00	0.04	0.06
9062	0.59	0.34	0.45	0.81	0.54	0.75	0.51	0.41	0.57	0.59	0.61	0.55	0.51	0.29	0.35	0.64	0.56	0.48	0.50
9068	0.97	0.29	0.31	0.69	0.40	0.92	0.97	0.41	0.63	0.90	0.78	0.61	0.70	0.80	0.43	0.42	0.95	0.24	0.47
9073	0.10	0.54	0.77	0.76	0.77	0.77	0.07	0.44	0.76	0.40	0.76	0.69	0.35	0.54	0.70	0.58	0.08	0.63	0.88
9081	0.67	0.33	0.05		0.03		0.49	0.07	0.02		0.01	0.03	0.02	0.02	0.01	0.08	0.88	0.06	0.01
9083	0.89	0.93	0.96	0.84	0.94	0.96	0.87	0.96	0.64	0.98	0.96	0.91	0.77	0.86	0.90	0.92	0.90	0.94	0.71
9087	0.01	0.04	0.06	0.68	0.07	0.11	0.01	0.02	0.35	0.01	0.41	0.23	0.36	0.15	0.15	0.15	0.00	0.16	0.62
9110	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	0.94	1.00	0.99	1.00	1.00	1.00	0.97	1.00	1.00	1.00
9111	0.18	0.10	0.23	0.08	0.21	0.02	0.34	0.10	0.12	0.48	0.13	0.21	0.28	0.16	0.15	0.21	0.12	0.44	0.24
9119	0.07	0.02	0.00	0.01	0.01	0.04	0.02	0.01	0.00	0.05	0.00	0.01	0.05	0.02	0.01	0.01	0.01	0.00	0.00
9128	0.82	0.87	0.98	0.62	0.95	0.51	0.70	0.72	0.81	0.79	0.78	0.54	0.56	0.56		0.97	0.95	0.39	0.51
9131	0.03	0.36	0.46	0.58	0.45	0.24	0.06	0.21	0.86	0.31	0.58	0.53	0.27	0.14	0.42	0.39	0.01	0.49	0.83

Sentence Source	wiley.com	gov.au	abc.net.au	nationalgeographic.com	smh.com.au	oreilly.com	springer.com	edu	wikipedia.org	nejm.org	bbc.com	time.com	mit.edu	au.news.yahoo.com	skynews.com.au	economist.com	ncbi.nlm.nih.gov	nytimes.com	imdb.com
9136	0.81	0.99	0.91	0.95	0.94	0.92	0.72	0.96	1.00	0.84	0.94	0.88	0.92	0.76	0.90	0.80	0.88	0.96	0.99
9139	0.76	0.43	0.22	0.17	0.49	0.31	0.85	0.45	0.05	0.25	0.39	0.37	0.50	0.66	0.28	0.40	0.98	0.41	0.42
9146	0.88	0.81	0.86	0.89	0.97	0.84	0.88	0.82	0.49	0.94	0.97	0.95	0.84	0.98	0.97	0.90	0.96	0.93	0.88
9147	0.02	0.05	0.02	0.00	0.38		0.04	0.01	0.08	0.27	0.09	0.35	0.01	0.66		0.00	0.06	0.16	0.40
9148	0.05	0.26	0.71	0.19	0.67		0.09	0.18	0.08	0.03	0.56	0.40	0.67	0.78		0.61	0.06	0.33	0.73
9161	0.00	0.02	0.02	0.02	0.13	0.01	0.01	0.00	0.00	0.00	0.00	0.03	0.17	0.11	0.04	0.14	0.00	0.00	0.00
9163	0.99	0.96	0.82	0.96	0.97	0.75	0.99	0.97	1.00		0.98	0.90	0.89	0.98	0.92	0.95	0.97	0.97	0.97
9167	0.01	0.04	0.06	0.08	0.24	0.09	0.01	0.01	0.01	0.29	0.15	0.27	0.04	0.09	0.24	0.28	0.00	0.19	0.43
9168	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.02	0.00	0.02	0.01	0.08	0.00	0.00	0.00
9171	0.98	1.00	0.97	0.63	0.88	0.99	0.99	0.99	1.00	0.88	0.96	0.97	0.96	1.00	0.95	0.97	0.99	0.96	1.00
9172	0.12	0.24	0.03	0.05	0.03	0.23	0.14	0.05	0.02	0.04	0.02	0.03	0.36	0.06	0.07	0.05	0.03	0.04	0.14
9179	0.57	0.50	0.22	0.77	0.11	0.87	0.63	0.80	0.83	0.14	0.56	0.38	0.81	0.32		0.07	0.24	0.33	0.61
9181	0.60	0.79	0.88	0.98		0.37	0.70	0.78	0.57	0.78			0.65			0.99	0.87	0.10	0.47
9190	0.16	0.14	0.91	0.59	0.87	0.08	0.11	0.10	0.84	0.34	0.96	0.68	0.15	0.83	0.83	0.85	0.23	0.73	0.60
9213	0.70	0.97	0.97	0.85	0.94	0.95	0.75	0.98	0.95	0.42	0.98	0.93	0.90	0.83	0.98	0.84	0.82	0.95	0.98
9221	0.99	1.00	0.77	0.93	0.94	0.95	0.96	0.99	0.85	1.00	0.89	0.87	0.90	0.89	0.88	0.94	1.00	0.86	0.86
9224	1.00	1.00	1.00		1.00		1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
9226	0.06	0.06	0.09	0.19	0.05	0.07	0.04	0.03	0.01	0.04	0.11	0.03	0.08	0.10	0.03	0.05	0.02	0.05	0.09
9228	0.01	0.04	0.06	0.03	0.18	0.16	0.01	0.03	0.00	0.00	0.08	0.16	0.27	0.25	0.19	0.32	0.00	0.02	0.01
9234	1.00	1.00	0.96	0.94	0.98	1.00	1.00	1.00	0.99	1.00	0.99	0.98	0.98			0.93	1.00	0.99	0.98
9259	0.00	0.00	0.05	0.02	0.04	0.00	0.00	0.00	0.00	0.00	0.04	0.09	0.01	0.03	0.07	0.08	0.00	0.08	0.29
9280	0.88	0.19	0.22	0.26	0.36	0.38	0.80	0.51	0.07	0.43	0.14	0.53	0.49	0.25	0.16	0.45	0.96	0.31	0.15
9289	0.44	0.30	0.31	0.21	0.31	0.14	0.54	0.30	0.39	0.15	0.04	0.20	0.28	0.21	0.34	0.38	0.31	0.37	0.30
9309	0.19	0.06	0.07	0.18	0.14	0.10	0.21	0.09	0.09	0.46	0.18	0.23	0.27	0.31	0.05	0.03	0.20	0.13	0.21
9311	0.66	0.56	0.28	0.31	0.34	0.52	0.58	0.42	0.60	0.45	0.08	0.33	0.43	0.31	0.28	0.44	0.65	0.25	0.14
9319	0.98	1.00	1.00		1.00	0.83	0.98	0.97	0.94	1.00	0.99	1.00	0.89	0.98	1.00	0.99	0.99	0.99	0.99
9323	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.89	0.99	0.98	0.98	0.99	0.99	1.00	0.99	1.00
9328	0.42	0.69	0.74	0.56	0.64	0.80	0.47	0.31	0.76	0.40	0.70	0.60	0.55	0.74		0.42	0.11	0.67	0.90
9342	0.56	0.89	0.91	0.90	0.70	0.71	0.53	0.80	0.85	0.28	0.77	0.69	0.82	0.71	0.67	0.58	0.39	0.67	0.72
9352	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00		0.90	1.00	1.00	1.00	1.00	1.00	1.00
9357	0.02	0.13	0.25	0.27	0.20	0.05	0.02	0.03	0.34	0.61	0.27	0.31	0.06	0.26	0.29	0.20	0.06	0.35	0.51
9366	0.92	0.77	0.24	0.57	0.49	0.43	0.91	0.80	0.94	0.87	0.44	0.37	0.64	0.70	0.31	0.34	1.00	0.48	0.94
9370	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.02	0.00	0.00	0.00
9372	0.58	0.76	0.72	0.72	0.83	0.86	0.62	0.73	0.70	0.58	0.59	0.67	0.46	0.58	0.81	0.75	0.07	0.65	0.72
9373	0.04	0.09	0.03	0.04	0.06	0.04	0.03	0.07	0.07	0.04	0.05	0.04	0.13	0.01	0.05	0.02	0.04	0.07	0.14
9378	1.00	1.00	1.00	0.96	0.99	1.00	1.00	1.00	0.99	0.99	1.00	0.99	0.99	1.00	1.00	0.99	1.00	0.99	0.93
9379	0.71	0.69	0.76	0.65	0.56	0.68	0.68	0.80	0.44	0.19	0.53	0.60	0.26	0.98	0.76	0.52	0.68	0.51	0.33
9394	0.02	0.10	0.68	0.27	0.26	0.03	0.03	0.03	0.30	0.02	0.54	0.17	0.21	0.46	0.24	0.17	0.06	0.35	0.44
9403	0.01	0.12	0.38	0.30	0.13	0.06	0.01	0.02	0.16	0.04	0.24	0.23	0.07	0.80	0.18	0.18	0.00	0.20	0.27
9415	0.02	0.34	0.07	0.08	0.17	0.35	0.04	0.10	0.01	0.11	0.05	0.12	0.28	0.04	0.19	0.24	0.00	0.03	0.02
9427	0.44	0.08	0.05	0.29	0.16	0.43	0.52	0.10	0.59	0.53	0.24	0.28	0.31	0.22	0.19	0.27	0.39	0.44	0.86
9434	0.80	0.65	0.71	0.27	0.83	0.87	0.73	0.77	0.80	0.94	0.88	0.84	0.84	0.89			0.85	0.74	0.89
9438	0.50	0.92	0.98	0.98	0.93	0.89	0.65	0.90	0.91	0.19	0.83	0.89	0.88	0.84		0.98	0.30	0.93	0.88
9452	0.31	0.71	0.97	0.86	0.92	0.81	0.36	0.63	1.00	0.38	0.93	0.92	0.70	0.73	0.95	0.87	0.25	0.93	0.84
9464	0.27	0.94	0.96	0.94	0.96	0.47	0.28	0.66	0.91	0.40	0.96	0.94	0.84	0.87		0.93	0.31	0.96	0.99
9467	0.83	0.62	0.76	0.91	0.83	0.73	0.71	0.72	0.97	0.85	0.89	0.50	0.77	0.48	0.75	0.87	0.81	0.86	0.92
9469	0.80	0.42	0.06	0.59	0.31	0.48	0.90	0.75	0.50	0.88	0.22	0.57	0.59	0.51	0.45	0.61	0.99	0.48	0.36
9482	0.08	0.06	0.10	0.11	0.08	0.08	0.02	0.02	0.00	0.06	0.00	0.06	0.12	0.03	0.08	0.09	0.01	0.04	0.01
9489	0.81	0.90	0.92		0.99		0.77	0.94	0.85	0.73	0.98	0.97	0.93	0.98			0.89		0.99
9490	0.03	0.31	0.22	0.29	0.23	0.40	0.03	0.23	0.41	0.15	0.14	0.32	0.36	0.10	0.15	0.59	0.00	0.30	0.36
9494	0.65	0.31	0.52	0.35	0.33	0.07	0.45	0.41	0.52	0.83	0.35	0.53	0.21	0.32	0.46	0.54	0.83	0.46	0.73
9508	0.28	0.05	0.01	0.02	0.04	0.19	0.41	0.06	0.00	0.30	0.00	0.04	0.24	0.10	0.02	0.09	0.37	0.00	0.01
9515	0.40	0.79	0.89	0.84	0.82	0.57	0.41	0.73	0.86	0.60	0.90	0.86	0.63	0.42	0.98	0.59	0.45	0.87	0.87
9519	0.88	0.74	0.88	0.70	0.90	0.93	0.84	0.96	0.85	0.72	0.88	0.93	0.75	0.98	0.75	0.72	0.99	0.90	0.97
9524	0.99	0.98					0.99	0.99	0.98	0.96			0.99	1.00			1.00	1.00	
9542	0.02	0.12	0.05	0.15	0.17	0.13	0.02	0.04	0.10	0.20	0.20	0.11	0.18	0.09	0.12	0.22	0.00	0.13	0.06
9556	0.35	0.30	0.39	0.36	0.56	0.62	0.46	0.47	0.19	0.49	0.10	0.46	0.51	0.56	0.29	0.58	0.38	0.35	0.82
9559	0.73	0.03	0.40	0.43	0.23	0.45	0.64	0.38	0.59	0.54	0.16	0.25	0.81	0.27	0.22	0.28	0.81	0.11	0.31
9573	0.35	0.17	0.41	0.73	0.67	0.31	0.28	0.11	0.02	0.41	0.75	0.38	0.30	0.84	0.39	0.34	0.42	0.33	0.36
9576	0.13	0.70	0.88	0.91	0.76	0.92	0.20	0.84	1.00	0.09	0.81	0.86	0.84	0.87	0.82	0.66	0.05	0.64	1.00
9582	0.97	0.85	0.87	0.86	0.85	0.59	0.93	0.95	0.84	0.99	0.83	0.79	0.48	0.94	0.50	0.65	0.99	0.58	0.82
9589	0.98	0.95	0.99	0.99		0.26	0.98	0.93	0.87	0.94	0.96		0.66	0.64		0.95	0.99	0.98	0.94
9597	0.42	0.70	0.91	0.63	0.72	0.64	0.39	0.52	0.53	0.75	0.72	0.66	0.70	0.80	0.71	0.66	0.34	0.57	0.81
9601	0.37	0.12	0.20	0.39	0.25	0.56	0.40	0.05	0.61	0.14	0.45	0.56	0.40	0.69	0.54	0.56	0.19	0.56	0.65
9618	0.99	1.00	0.99	0.99	0.98	1.00	0.99	1.00	1.00	0.91	0.99	0.99	0.97	0.99	0.99	1.00	1.00	0.97	0.92
9624	1.00	0.97	0.44	0.94	0.93	0.98	1.00	1.00	0.91	0.99	0.88	0.88	0.95	0.95	0.82	0.87	1.00	0.89	0.79

Sentence Source	wiley.com	gov.au	abc.net.au	nationalgeographic.com	smh.com.au	oreilly.com	springer.com	edu	wikipedia.org	nejm.org	bbc.com	time.com	mit.edu	au.news.yahoo.com	skynews.com.au	economist.com	ncbi.nlm.nih.gov	nytimes.com	imdb.com
9647	0.99	0.94	0.91	0.96	0.82	0.93	0.99	0.99	1.00	0.95	0.96	0.71	0.92	0.55	0.93	0.58	1.00	0.86	0.95
9648	0.04	0.69	0.69	0.37	0.62	0.09	0.04	0.56	0.37	0.08	0.15	0.40	0.28	0.21	0.56	0.50	0.06	0.82	0.37
9659	0.14	0.36	0.26	0.21	0.22	0.28	0.22	0.13	0.15	0.12	0.15	0.44	0.28	0.67	0.14	0.71	0.04	0.25	0.18
9673	0.46	0.34	0.68	0.56	0.55	0.49	0.49	0.52	0.56	0.52	0.79	0.58	0.64	0.79	0.65		0.28	0.42	0.91
9682	0.34	0.45	0.78		0.75	0.36	0.40	0.36	0.68	0.39	0.88	0.64	0.63	0.80	0.87	0.39	0.41	0.61	0.93
9684	0.09	0.55	0.62	0.31	0.80	0.89	0.49	0.58	0.69	0.54	0.99	0.78	0.74	0.73	0.97	0.75	0.04	0.77	1.00
9688	0.51	0.78	0.60	0.79	0.47	0.80	0.45	0.63	0.55	0.71	0.04	0.63	0.84	0.04	0.47	0.44	0.73	0.44	0.09
9689	0.94	0.95	1.00	0.88	1.00	0.25	0.79	0.79	0.78		1.00	1.00	0.16	0.98	0.99	0.99	0.95	0.99	0.99
9690	0.09	0.15	0.14	0.16	0.34	0.09	0.14	0.15	0.00	0.36	0.20	0.08	0.27	0.20	0.20	0.17	0.09	0.09	0.10
9695	0.97	0.96	0.98	0.87	0.95	0.92	0.96	0.97	1.00	0.85	0.95	0.92	0.87	0.83	0.96	0.85	1.00	0.92	0.99
9698	0.59	0.28	0.36	0.20	0.18	0.03	0.38	0.35	0.11	0.62	0.16	0.26	0.19	0.92	0.28	0.43	0.76	0.09	0.01
9699	0.59	0.75	0.98	0.67	0.99	0.96	0.67	0.95	0.85	0.96	0.96	0.96	0.71	0.95	0.97	0.97	0.85	0.92	0.88
9703	0.91	0.91	0.78	0.70	0.75		0.84	0.85	0.33	0.99	0.81	0.68	0.89	0.87	0.68	0.69	0.97	0.52	0.64
9708	0.93	0.92	0.90	0.84	0.83	0.75	0.84	0.93	0.94	0.90	0.72	0.84	0.78	0.87	0.84	0.81	0.93	0.88	0.95
9719	0.00	0.00	0.04	0.02	0.07	0.01	0.00	0.00	0.00	0.00	0.04	0.02	0.01	0.02	0.01	0.04	0.00	0.02	0.01
9720	0.31	0.22	0.47	0.30	0.43	0.08	0.23	0.15	0.05	0.33	0.34	0.24	0.24	0.39	0.43	0.24	0.19	0.22	0.15
9722	0.28	0.07	0.10	0.39	0.31	0.12	0.30	0.38	0.22	0.61	0.24	0.51	0.35	0.69	0.56	0.41	0.58	0.45	0.59
9725	0.68	0.96	0.99	0.90	0.99	0.90	0.77	0.99	0.99	0.87	0.99	0.99	0.79	0.95	1.00	0.95	0.83	0.96	0.98
9727	0.27	0.83	0.84	0.79	0.71	0.84	0.22	0.39	0.86	0.21	0.77	0.53	0.62	0.67	0.94	0.61	0.22	0.59	0.79
9731	0.04	0.14	0.14	0.03	0.22	0.19	0.03	0.04	0.03	0.00	0.23	0.21	0.70	0.50	0.21	0.18	0.02	0.21	0.37
9732	0.39	0.20	0.29	0.07	0.38	0.01	0.64	0.03	0.01	0.07	0.28	0.07	0.32	0.05	0.17	0.28	0.05	0.02	0.12
9739	0.83	0.18	0.28	0.12	0.15	0.28	0.78	0.31	0.00	0.72	0.03	0.08	0.22	0.22	0.15	0.26	0.81	0.07	0.03
9740	0.99	0.61	0.35	0.38	0.29	0.61	0.98	0.85	0.00	0.82	0.15	0.29	0.39	0.65	0.25	0.26	1.00	0.25	0.07
9758	0.20	0.04	0.17	0.06	0.27	0.07	0.28	0.08	0.00	0.52	0.12	0.18	0.04	0.62	0.24	0.15	0.44	0.08	0.03
9769	0.63	0.84	0.85	0.71	0.75	0.88	0.49	0.62	0.93	0.63	0.85	0.78	0.73	0.74	0.87	0.70	0.37	0.65	0.86
9780	0.96	0.98	0.91	0.99	0.97	0.98	0.99	0.99	0.96	0.84	0.97	0.96	0.98	0.87	1.00	1.00	0.92	0.99	0.99
9791	0.13	0.25	0.29	0.19	0.24		0.18	0.25	0.77	0.24	0.68	0.30	0.39	0.11	0.32	0.17	0.14	0.38	0.42
9797	0.37	0.19	0.06	0.21	0.13	0.17	0.14	0.28	0.00	0.47	0.01	0.07	0.11	0.09	0.13	0.04	0.83	0.04	0.02
9812	0.99	0.99	0.99	0.95	0.97	0.96	0.99	0.99	1.00	1.00	0.99	0.90	0.97	0.98	0.99	0.88	1.00	0.94	0.98
9819	0.28	0.19	0.53	0.50	0.53	0.03	0.35	0.10	0.00	0.39	0.74	0.17	0.14	0.24	0.38	0.16	0.46	0.17	0.05
9829	0.43	0.04	0.02	0.02	0.01	0.01	0.36	0.09	0.06	0.12	0.01	0.01	0.30	0.03	0.01	0.03	0.37	0.02	0.00
9831	0.99	0.97	0.81	0.95	0.96	0.82	0.97	0.92	0.88	0.94	0.98	0.95	0.78	0.96	0.98	0.90	1.00	0.91	0.84
9848	0.50	0.74	0.77	0.41	0.59	0.75	0.43	0.23	0.22	0.64	0.55	0.45	0.26	0.41	0.47	0.61	0.50	0.31	0.17
9861	0.31	0.16	0.17	0.35	0.23	0.58	0.24	0.45	0.65	0.39	0.33	0.47	0.38	0.10	0.35	0.20	0.10	0.40	0.79
9868	0.96	1.00	0.97	0.99	0.96	0.98	0.95	1.00	1.00	0.99	0.98	0.97	0.98	0.90	0.97	0.96	0.93	0.98	1.00
9876	0.20	0.04	0.68	0.38	0.61	0.10	0.08	0.16	0.29	0.58	0.44	0.47	0.19	0.28	0.55	0.38	0.27	0.40	0.23
9899	0.34	0.20	0.30	0.12	0.48	0.06	0.34	0.07	0.17	0.62	0.48	0.19	0.15	0.47	0.25	0.35	0.42	0.35	0.15
9901	0.18	0.18	0.41	0.47	0.35	0.11	0.40	0.14	0.00	0.60	0.46	0.28	0.25	0.57	0.53	0.26	0.77	0.10	0.00
9902	0.48	0.87	0.39	0.89	0.57	0.67	0.86	0.61	0.64	0.61	0.81	0.78	0.57	0.56	0.83	0.68	0.74	0.62	0.69
9919	0.91	0.15	0.36	0.15	0.16	0.35	0.93	0.67	0.18	0.86	0.23	0.23	0.47	0.37	0.20	0.27	0.96	0.15	0.08
9956	0.00	0.08	0.03	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.07	0.01	0.03	0.01	0.05	0.03	0.00	0.00	0.00
9965	0.95	0.92	0.66	0.97	0.84	0.94	0.84	0.67	0.98	0.87	0.86	0.95	0.68	1.00	0.93	0.82	1.00	0.85	0.64
9967	0.95	0.95	0.81	0.66	0.84	0.21	0.91	0.73	0.55	0.90	0.96	0.55	0.53	0.94	0.29	0.84	0.95	0.30	0.43
9974	0.81	0.73	0.83	0.64	0.76	0.57	0.82	0.91	0.49	0.95	0.77	0.76	0.63	0.61	0.90	0.72	0.89	0.84	0.70
9975	0.17	0.10	0.21	0.04	0.19	0.03	0.22	0.13	0.11	0.54	0.12	0.21	0.23	0.35	0.18	0.21	0.10	0.35	0.25
9979	0.01	0.02	0.36	0.02	0.66	0.03	0.01	0.01	0.00	0.06	0.38	0.44	0.11	0.35	0.52	0.63	0.00	0.40	0.22
9985	0.73	0.95	0.99	1.00	0.89	0.96	0.88	0.98	0.99	0.84	0.91	0.94	0.91	0.81	0.94	0.93	0.15	0.90	0.92
10003	0.97	0.92	0.96	0.97	0.91	0.96	0.97	0.98	0.97		0.99	0.90	0.99	0.97	0.96	0.73	1.00	0.92	0.97
10009	0.84	0.75	0.33	0.75	0.64	0.85	0.80	0.78	0.37	0.96	0.33	0.58	0.68	0.80	0.45	0.63	0.97	0.61	0.15
10012	0.98	1.00	0.65	0.84	0.91	0.98	0.99	0.99	0.89		0.87	0.89	0.96	0.82	0.90	0.86	1.00	0.90	0.93
10038	0.08	0.14	0.04	0.12	0.15	0.10	0.10	0.09	0.07	0.76	0.08	0.08	0.19	0.25	0.07	0.20	0.02	0.11	0.02
10040	0.59	0.04	0.08	0.20	0.14	0.06	0.51	0.12	0.00	0.62	0.03	0.08	0.30	0.06	0.21	0.16	0.19	0.09	0.03
10041	0.30	0.91	0.61	0.42	0.88	0.53	0.33	0.74	0.48	0.61	0.39	0.77	0.41	0.68	0.86	0.65	0.48	0.86	0.87
10047	0.06	0.12	0.36	0.05	0.28	0.08	0.05	0.48	0.25	0.33	0.31	0.67	0.50	0.30	0.26	0.55	0.01	0.61	0.74
10054	0.20	0.50	0.78	0.81	0.76	0.77	0.31	0.43	0.56	0.32	0.84	0.76	0.50	0.81	0.90	0.70	0.25	0.66	0.73
10059	1.00	1.00	1.00	0.99	1.00	0.97	1.00	1.00	1.00		1.00	1.00	0.97	1.00		0.99	1.00	1.00	0.99
10064	0.01	0.21	0.50	0.56	0.64	0.19	0.01	0.03	0.38	0.03	0.49	0.48	0.10	0.45	0.36	0.21	0.01	0.27	0.69
10086	0.73	0.96	0.91	0.91	0.93	0.99	0.70	0.86	0.86	0.28	0.97	0.86	0.90	0.98	0.96	0.90	0.58	0.83	0.90
10090	0.99	1.00	0.99		0.97		0.99	0.99	0.99	0.91	0.99	0.98	0.99	0.98		0.99	0.96	0.98	0.93
10102	0.95	0.54	0.39	0.88	0.76	0.81	0.96	0.94	0.71	0.83	0.40	0.49	0.77	0.33	0.88	0.49	0.99	0.33	0.24
10118	0.30	0.61	0.97	0.27	0.82	0.94	0.32	0.95	1.00	0.15	0.90	0.89	0.80	0.87	0.84	0.75	0.02	0.89	0.98
10134	0.45	0.70	0.66		0.87		0.31	0.63	0.85	0.65	0.92	0.83	0.22	0.90			0.37	0.83	0.97
10156	0.64	0.08	0.14	0.18	0.07	0.04	0.51	0.17	0.15	0.26	0.06	0.11	0.20	0.16	0.04	0.06	0.63	0.08	0.01
10177	0.76	0.88	0.77	0.42	0.85	0.48	0.68	0.87	0.68	0.71	0.67	0.79	0.52	0.78	0.81	0.64	0.65	0.77	0.32
10182	0.49	0.93	0.83	0.69	0.78	0.79	0.52	0.85	1.00	0.43	0.97	0.78	0.69	0.45	0.74	0.79	0.40	0.86	0.95
10209	0.83	0.99	1.00	0.97	1.00		0.88	0.97	1.00	1.00	1.00	1.00	0.90	1.00	1.00	0.99	0.94	1.00	1.00
10228	0.09	0.47	0.77	0.16	0.53	0.52	0.15	0.18	0.00	0.37	0.40	0.19	0.30	0.34	0.30	0.37	0.15	0.29	0.01

Sentence Source	wiley.com	gov.au	abc.net.au	nationalgeographic.com	smh.com.au	oreilly.com	springer.com	edu	wikipedia.org	nejm.org	bbc.com	time.com	mit.edu	au.news.yahoo.com	skynews.com.au	economist.com	ncbi.nlm.nih.gov	nytimes.com	imdb.com
10230	0.56	0.97	0.89	0.69	0.68	0.92	0.81	0.70	0.54	0.86	0.85	0.74	0.71	0.81	0.85	0.83	0.87	0.71	0.62
10246	0.97	0.86	0.82	0.65	0.66	0.84	0.99	0.87	0.64	0.91	0.36	0.55	0.85	0.72	0.66	0.68	1.00	0.75	0.00
10254	0.43	0.47	0.59	0.45	0.55	0.30	0.39	0.77	0.30	0.69	0.42	0.36	0.15	0.21	0.58	0.33	0.49	0.28	0.95
10257	0.81	0.91	0.95	0.62	0.95	0.99	0.90	0.95	0.66	0.88	0.94	0.88	0.76	0.96	0.93	0.94	0.89	0.87	0.99
10258	0.99	0.79	0.80	0.77	0.79	0.70	0.99	0.95	0.19	0.98	0.76	0.71	0.68	0.74	0.56	0.53	1.00	0.68	0.40
10267	0.93	0.93	0.73	0.90	0.96	0.81	0.94	0.99	0.99	1.00	0.97	0.89	0.89	0.97	0.94	0.96	0.98	0.96	0.96
10275	0.28	0.43	0.51	0.60	0.41	0.49	0.32	0.48	0.27	0.04	0.49	0.36	0.57	0.93	0.35	0.48	0.03	0.28	0.26
10279	0.03	0.30	0.10	0.01	0.16	0.01	0.05	0.06	0.21	0.13	0.02	0.06	0.12	0.01	0.02	0.10	0.01	0.09	0.02
10281	0.35	0.06	0.06	0.06	0.03	0.15	0.30	0.02	0.05	0.14	0.02	0.06	0.05	0.06	0.02	0.08	0.47	0.06	0.04
10285	0.71	0.92	0.77	0.16	0.64	0.42	0.72	0.70	0.84	0.24	0.40	0.53	0.60	0.09	0.63	0.66	0.56	0.61	0.43
10296	0.99	0.93	0.95	0.97	0.94	0.97	0.98	0.99	0.87	0.99	0.93	0.95	0.91	0.99	0.94	0.96	1.00	0.92	0.85
10335	0.99	0.97	0.88	0.72	0.92	0.90	0.98	0.92	0.90	0.97	0.88	0.85	0.68	0.91	0.86	0.85	1.00	0.81	0.73
10344	0.01	0.02	0.01	0.01	0.01	0.00	0.01	0.01	0.04	0.09	0.00	0.01	0.02	0.00	0.00	0.01	0.00	0.01	0.01
10350	0.33	0.09	0.22	0.25	0.33	0.42	0.24	0.06	0.04	0.11	0.35	0.28	0.39	0.25	0.24	0.34	0.13	0.28	0.75
10373	0.81	0.06	0.01	0.13	0.02	0.11	0.61	0.26	0.00	0.35	0.01	0.02	0.18	0.04	0.02	0.11	0.56	0.01	0.03
10379	0.96	0.99	0.99	0.99	1.00	1.00	0.95	1.00	1.00	0.97	1.00	0.98	0.97	1.00	1.00	0.98	1.00	1.00	1.00
10382	0.33	0.77	0.86	0.86	0.84	0.65	0.59	0.60	0.65	0.69	0.86	0.75	0.53	0.90	0.93	0.81	0.44	0.89	0.94
10383	0.55	0.65	0.70	0.60	0.36	0.12	0.48	0.29	0.56	0.83	0.70	0.52	0.23	0.30	0.84	0.14	0.88	0.35	0.83
10396	0.28	0.36	0.92	0.46	0.62	0.24	0.13	0.09	0.01	0.23	0.76	0.63	0.24	0.68	0.80	0.41	0.10	0.49	0.33
10398	0.82	0.77	0.87		0.73	0.56	0.72	0.86	0.35	0.99	0.87	0.92	0.66	0.65		0.93	0.97	0.84	0.39
10417	0.09	0.21	0.15	0.15	0.27	0.54	0.16	0.17	0.00	0.11	0.07	0.38	0.31	0.88	0.30	0.62	0.01	0.03	0.06
10419	0.50	0.93	0.97	0.87	0.85	0.97	0.60	0.68	1.00	0.77	0.98	0.87	0.89	0.94	0.96	0.74	0.79	0.98	1.00
10425	0.49	0.67	0.74	0.56	0.86	0.43	0.52	0.73	0.28	0.87	0.69	0.71	0.33	0.85	0.72	0.55	0.80	0.68	0.63
10426	0.86	0.92	0.99				0.76	0.89	0.98	0.98	1.00	1.00	0.90	1.00		0.98	0.85	1.00	1.00
10434	1.00	0.53	0.78	0.88	0.36	0.86	1.00	0.99	0.91		0.74	0.29	0.96	0.62	0.55	0.34	1.00	0.19	0.37
10466	0.02	0.22	0.49	0.55	0.65	0.16	0.01	0.03	0.36	0.04	0.52	0.51	0.10	0.45	0.36	0.32	0.01	0.24	0.55
10468	0.28	0.10	0.03	0.06	0.06	0.28	0.18	0.12	0.26	0.03	0.05	0.13	0.12	0.04	0.05	0.26	0.29	0.24	0.05
10481	0.83	0.78	0.36	0.73	0.58	0.25	0.84	0.71	0.13	0.80	0.38	0.47	0.67	0.78	0.51	0.57	0.99	0.72	0.52
10493	0.57	0.91	0.98	0.98	0.93	0.91	0.62	0.91	0.93	0.36	0.97	0.90	0.96	0.92	0.98	0.97	0.45	0.94	0.95
10500	0.11	0.63	0.40	0.56	0.72	0.53	0.31	0.23	0.34	0.57	0.56	0.51	0.58	0.56	0.66	0.72	0.04	0.46	0.50
10501	1.00	0.99	0.83	0.81	0.91	0.97	1.00	0.99	0.95	1.00	0.90	0.93	0.96	0.96	0.96	0.88	1.00	0.93	0.90
10504	0.26	0.97	0.93	0.87	0.93	0.93	0.64	0.98	1.00	0.95	0.94	0.93	0.91	0.99	0.98	0.75	0.44	0.94	0.90
10505	0.58	0.14	0.22	0.39	0.05	0.32	0.29	0.41	0.40	0.44	0.04	0.20	0.50	0.54	0.07	0.27	0.57	0.40	0.05
10523	0.15	0.01	0.03	0.01	0.01	0.01	0.14	0.03	0.01	0.06	0.03	0.00	0.19	0.01	0.01	0.02	0.01	0.01	0.00
10524	0.16	0.36	0.35	0.53	0.80	0.84	0.31	0.39	0.36	0.41	0.53	0.61	0.61	0.59	0.67	0.87	0.04	0.66	0.43
10528	0.46	0.19	0.03	0.15	0.11	0.31	0.23	0.12	0.00	0.22	0.01	0.17	0.15	0.17	0.05	0.04	0.86	0.03	0.01
10544	0.78	0.71	0.36	0.87	0.76	0.89	0.78	0.92	0.81	0.93	0.60	0.64	0.88	0.66	0.52	0.59	0.91	0.75	0.89
10545	0.04	0.27	0.51	0.44	0.09	0.47	0.14	0.12	0.36	0.27	0.21	0.54	0.22	0.39		0.23	0.02	0.38	0.32
10551	0.75	0.13	0.05	0.18	0.17	0.60	0.58	0.54	0.11	0.92	0.05	0.20	0.41	0.38	0.20	0.31	0.76	0.25	0.16
10552	0.97	0.94	0.97	0.99	0.96	0.99	0.97	0.98	0.99	0.46	0.99	0.96	0.98	0.96	0.99	0.98	0.96	0.96	0.98
10577	0.40	0.05	0.76	0.17	0.07	0.21	0.35	0.06	0.25	0.43	0.25	0.10	0.13	0.09	0.05	0.05	0.28	0.08	0.15
10578	0.90	0.93	0.80	0.91	0.76	0.60	0.97	0.90	0.87	0.98	0.69	0.87	0.71	0.65	0.72	0.72	0.99	0.60	0.58
10584	0.04	0.01	0.02	0.07	0.04	0.15	0.10	0.00	0.01	0.01	0.02	0.01	0.24	0.02	0.01	0.04	0.06	0.03	0.05
10597	0.11	0.31	0.91	0.73	0.89	0.80	0.12	0.77	0.65	0.18	0.98	0.91	0.59	0.96	0.90	0.95	0.02	0.83	0.88
10599	0.35	0.54	0.32	0.23	0.49	0.08	0.24	0.25	0.24	0.65	0.41	0.23	0.19	0.40	0.40	0.20	0.41	0.26	0.04
10605	0.30	0.53	0.71	0.15	0.39	0.84	0.19	0.58	0.70	0.15	0.42	0.58	0.67	0.54	0.58	0.76	0.04	0.46	0.52
10606	0.78	0.21	0.23	0.24	0.28	0.38	0.80	0.41	0.04	0.25	0.34	0.27	0.18	0.62	0.29	0.40	0.37	0.19	0.16
10614	0.71	0.73	0.91		0.98	0.85	0.95	0.85	0.84	0.83	0.98	0.93	0.72	0.96	0.98	0.96	0.96	0.97	0.98
10622	0.01	0.08	0.12	0.08	0.19	0.18	0.01	0.09	0.00	0.01	0.06	0.23	0.59	0.03	0.29	0.35	0.00	0.02	0.03
10627	0.94	0.67	0.80	0.60	0.82	0.64	0.92	0.78	0.28	0.92	0.56	0.81	0.76	0.88	0.96	0.88	0.92	0.85	0.73
10629	0.23	0.35	0.62	0.59	0.92	0.85	0.14	0.15	0.41	0.15	0.78	0.57	0.36	0.33	0.87	0.73	0.02	0.60	0.92
10637	0.00	0.01	0.02	0.02	0.03	0.01	0.00	0.00	0.00	0.01	0.02	0.08	0.03	0.03	0.01	0.04	0.00	0.10	0.02
10639	0.54	0.91	0.94	0.92	0.95	0.93	0.80	0.86	1.00	0.68	0.99	0.96	0.87	0.75	0.99	0.97	0.42	0.92	1.00
10641	0.02	0.40	0.64	0.14	0.48	0.83	0.03	0.11	0.16	0.18	0.19	0.35	0.37	0.58	0.25	0.48	0.03	0.40	0.37
10642	0.46	0.34	0.19	0.04	0.25		0.49	0.30	0.33	0.61	0.36	0.08	0.36	0.72	0.32	0.06	0.52	0.16	0.06
10643	0.93	0.99	1.00	0.93	0.97	0.91	0.91	0.97	0.95	0.84	0.99	0.86	0.80	0.93	0.98	0.88	0.77	0.85	0.94
10645	0.03	0.56	0.87	0.70	0.97	0.22	0.02	0.14	0.26	0.15	0.80	0.83	0.20	1.00		0.95	0.01	0.62	0.08
10646	0.35	0.35	0.15	0.19	0.01	0.06	0.36	0.10	0.00	0.10	0.06	0.11	0.27	0.10	0.08	0.16	0.31	0.15	0.14
10650	0.81	0.45	0.29	0.48	0.07	0.47	0.71	0.52	0.51	0.72	0.39	0.42	0.54	0.25	0.32	0.39	0.63	0.01	0.02
10692	1.00	1.00	1.00	0.98	1.00	1.00	1.00	1.00	1.00	0.93	1.00	0.99	1.00	1.00	0.99	0.99	0.99	0.99	0.98
10703	0.04	0.18	0.04	0.07	0.07	0.35	0.35	0.10	0.00	0.33	0.07	0.10	0.22	0.21	0.04	0.06	0.72	0.03	0.02
10704	0.99	0.99	1.00	0.99	0.99	0.98	0.99	0.99	1.00	0.98	0.99	0.99	0.97	1.00	0.99	0.99	1.00	0.98	0.98

B.2 Results

All the results from Section 5.3 were extracted from raw figures shown below.

B.2.1 1(a) Data Augmentation — Bayesian Framework

The subsequent tables illustrate how the performance changes as the PPC range varies. "+" in Bayes Factor indicates that BF is activated (See Section 5.1.5).

Range	Lower	Upper	Precision	Recall	F1	Accuracy	Coverage	Bayes Factor
50% ± 1%	49%	51%	83.33%	100.00%	90.91%	94.29%	10.67%	+
50% ± 2%	48%	52%	80.00%	85.71%	82.76%	87.50%	12.20%	+
50% ± 3%	47%	53%	82.35%	82.35%	82.35%	86.36%	13.41%	+
50% ± 4%	46%	54%	82.35%	82.35%	82.35%	86.67%	13.72%	+
50% ± 5%	45%	55%	82.35%	82.35%	82.35%	87.23%	14.33%	+
50% ± 6%	44%	56%	78.95%	78.95%	78.95%	84.62%	15.85%	+
50% ± 7%	43%	57%	78.95%	71.43%	75.00%	82.14%	17.07%	+
50% ± 8%	42%	58%	69.57%	69.57%	69.57%	78.12%	19.51%	+
50% ± 9%	41%	59%	69.57%	66.67%	68.09%	78.26%	21.04%	+
50% ± 10%	40%	60%	70.83%	60.71%	65.38%	76.00%	22.87%	+
50% ± 11%	39%	61%	68.00%	58.62%	62.96%	74.03%	23.48%	+
50% ± 12%	38%	62%	65.38%	56.67%	60.71%	72.84%	24.70%	+
50% ± 13%	37%	63%	65.52%	57.58%	61.29%	72.73%	26.83%	+
50% ± 14%	36%	64%	64.52%	58.82%	61.54%	72.53%	27.74%	+
50% ± 15%	35%	65%	66.67%	57.89%	61.97%	72.45%	29.88%	+
50% ± 16%	34%	66%	65.71%	54.76%	59.74%	70.75%	32.32%	+
50% ± 17%	33%	67%	63.89%	54.76%	58.97%	70.37%	32.93%	+
50% ± 18%	32%	68%	63.16%	54.55%	58.54%	69.37%	33.84%	+
50% ± 19%	31%	69%	64.10%	52.08%	57.47%	68.64%	35.98%	+
50% ± 20%	30%	70%	66.67%	51.85%	58.33%	68.25%	38.41%	+
50% ± 21%	29%	71%	65.91%	52.73%	58.59%	68.70%	39.94%	+
50% ± 22%	28%	72%	62.00%	54.39%	57.94%	67.39%	42.07%	+
50% ± 23%	27%	73%	62.75%	55.17%	58.72%	68.75%	43.90%	+
50% ± 24%	26%	74%	61.82%	55.74%	58.62%	68.21%	46.04%	+
50% ± 25%	25%	75%	60.34%	54.69%	57.38%	67.09%	48.17%	+
50% ± 26%	24%	76%	59.38%	55.88%	57.58%	66.47%	50.91%	+
50% ± 27%	23%	77%	57.35%	55.71%	56.52%	65.32%	52.74%	+
50% ± 28%	22%	78%	56.52%	54.93%	55.71%	64.57%	53.35%	+
50% ± 29%	21%	79%	56.52%	54.93%	55.71%	64.57%	53.35%	+
50% ± 30%	20%	80%	56.52%	54.17%	55.32%	64.61%	54.27%	+
50% ± 31%	19%	81%	54.17%	54.17%	54.17%	63.93%	55.79%	+
50% ± 32%	18%	82%	52.63%	54.05%	53.33%	62.96%	57.62%	+
50% ± 33%	17%	83%	52.56%	53.95%	53.25%	62.50%	58.54%	+
50% ± 34%	16%	84%	52.56%	53.95%	53.25%	62.69%	58.84%	+
50% ± 35%	15%	85%	51.90%	53.95%	52.90%	62.76%	59.76%	+
50% ± 36%	14%	86%	53.09%	55.13%	54.09%	63.32%	60.67%	+
50% ± 37%	13%	87%	52.44%	55.13%	53.75%	63.37%	61.59%	+
50% ± 38%	12%	88%	52.38%	55.00%	53.66%	63.29%	63.11%	+
50% ± 39%	11%	89%	52.38%	55.00%	53.66%	64.49%	65.24%	+
50% ± 40%	10%	90%	52.94%	55.56%	54.22%	64.81%	65.85%	+
50% ± 41%	9%	91%	53.49%	55.42%	54.44%	64.84%	66.77%	+
50% ± 42%	8%	92%	53.93%	55.81%	54.86%	64.57%	67.99%	+
50% ± 43%	7%	93%	52.75%	55.17%	53.93%	63.88%	69.21%	+
50% ± 44%	6%	94%	52.63%	56.18%	54.35%	63.64%	70.43%	+
50% ± 45%	5%	95%	51.00%	56.67%	53.68%	62.71%	71.95%	+
50% ± 46%	4%	96%	52.83%	58.95%	55.72%	63.37%	74.09%	+
50% ± 47%	3%	97%	50.45%	58.33%	54.11%	61.85%	75.91%	+
50% ± 48%	2%	98%	50.00%	58.33%	53.85%	61.60%	76.22%	+
50% ± 49%	1%	99%	49.14%	58.76%	53.52%	61.33%	78.05%	+
50% ± 50%	0%	100%	50.00%	57.52%	53.50%	60.35%	86.89%	+

Table B.1 How the performance changes when varying the PPC range (the 1^{st} run).

Range	Lower	Upper	Precision	Recall	F1	Accuracy	Coverage	Bayes Factor
50% ± 1%	49%	51%	83.33%	100.00%	90.91%	94.29%	10.67%	+
50% ± 2%	48%	52%	80.00%	85.71%	82.76%	87.50%	12.20%	+
50% ± 3%	47%	53%	82.35%	82.35%	82.35%	86.36%	13.41%	+
50% ± 4%	46%	54%	82.35%	82.35%	82.35%	86.67%	13.72%	+
50% ± 5%	45%	55%	82.35%	82.35%	82.35%	87.23%	14.33%	+
50% ± 6%	44%	56%	78.95%	78.95%	78.95%	84.62%	15.85%	+
50% ± 7%	43%	57%	78.95%	71.43%	75.00%	82.14%	17.07%	+
50% ± 8%	42%	58%	69.57%	69.57%	69.57%	78.12%	19.51%	+
50% ± 9%	41%	59%	69.57%	66.67%	68.09%	78.26%	21.04%	+
50% ± 10%	40%	60%	70.83%	60.71%	65.38%	76.00%	22.87%	+
50% ± 11%	39%	61%	68.00%	58.62%	62.96%	74.03%	23.48%	+
50% ± 12%	38%	62%	65.38%	56.67%	60.71%	72.84%	24.70%	+
50% ± 13%	37%	63%	65.52%	57.58%	61.29%	72.73%	26.83%	+
50% ± 14%	36%	64%	64.52%	58.82%	61.54%	72.53%	27.74%	+
50% ± 15%	35%	65%	66.67%	57.89%	61.97%	72.45%	29.88%	+
50% ± 16%	34%	66%	65.71%	54.76%	59.74%	70.75%	32.32%	+
50% ± 17%	33%	67%	63.89%	54.76%	58.97%	70.37%	32.93%	+
50% ± 18%	32%	68%	63.16%	54.55%	58.54%	69.37%	33.84%	+
50% ± 19%	31%	69%	64.10%	52.08%	57.47%	68.64%	35.98%	+
50% ± 20%	30%	70%	66.67%	51.85%	58.33%	68.25%	38.41%	+
50% ± 21%	29%	71%	65.91%	52.73%	58.59%	68.70%	39.94%	+
50% ± 22%	28%	72%	62.00%	54.39%	57.94%	67.39%	42.07%	+
50% ± 23%	27%	73%	62.75%	55.17%	58.72%	68.75%	43.90%	+
50% ± 24%	26%	74%	61.82%	55.74%	58.62%	68.21%	46.04%	+
50% ± 25%	25%	75%	60.34%	54.69%	57.38%	67.09%	48.17%	+
50% ± 26%	24%	76%	59.38%	55.88%	57.58%	66.47%	50.91%	+
50% ± 27%	23%	77%	57.35%	55.71%	56.52%	65.32%	52.74%	+
50% ± 28%	22%	78%	56.52%	54.93%	55.71%	64.57%	53.35%	+
50% ± 29%	21%	79%	56.52%	54.93%	55.71%	64.57%	53.35%	+
50% ± 30%	20%	80%	56.52%	54.17%	55.32%	64.61%	54.27%	+
50% ± 31%	19%	81%	54.17%	54.17%	54.17%	63.93%	55.79%	+
50% ± 32%	18%	82%	52.63%	54.05%	53.33%	62.96%	57.62%	+
50% ± 33%	17%	83%	52.56%	53.95%	53.25%	62.50%	58.54%	+
50% ± 34%	16%	84%	52.56%	53.95%	53.25%	62.69%	58.84%	+
50% ± 35%	15%	85%	51.90%	53.95%	52.90%	62.76%	59.76%	+
50% ± 36%	14%	86%	53.09%	55.13%	54.09%	63.32%	60.67%	+
50% ± 37%	13%	87%	52.44%	55.13%	53.75%	63.37%	61.59%	+
50% ± 38%	12%	88%	52.38%	55.00%	53.66%	63.29%	63.11%	+
50% ± 39%	11%	89%	52.38%	55.00%	53.66%	64.49%	65.24%	+
50% ± 40%	10%	90%	52.94%	55.56%	54.22%	64.81%	65.85%	+
50% ± 41%	9%	91%	53.49%	55.42%	54.44%	64.84%	66.77%	+
50% ± 42%	8%	92%	53.93%	55.81%	54.86%	64.57%	67.99%	+
50% ± 43%	7%	93%	52.75%	55.17%	53.93%	63.88%	69.21%	+
50% ± 44%	6%	94%	52.63%	56.18%	54.35%	63.64%	70.43%	+
50% ± 45%	5%	95%	51.00%	56.67%	53.68%	62.71%	71.95%	+
50% ± 46%	4%	96%	52.83%	58.95%	55.72%	63.37%	74.09%	+
50% ± 47%	3%	97%	50.45%	58.33%	54.11%	61.85%	75.91%	+
50% ± 48%	2%	98%	50.00%	58.33%	53.85%	61.60%	76.22%	+
50% ± 49%	1%	99%	49.14%	58.76%	53.52%	61.33%	78.05%	+
50% ± 50%	0%	100%	50.00%	57.52%	53.50%	60.35%	86.89%	+

Table B.2 How the performance changes when varying the PPC range (the 2^{nd} run).

Range	Lower	Upper	Precision	Recall	F1	Accuracy	Coverage	Bayes Factor
50% ± 1%	49%	51%	83.33%	100.00%	90.91%	94.29%	10.67%	+
50% ± 2%	48%	52%	80.00%	85.71%	82.76%	87.50%	12.20%	+
50% ± 3%	47%	53%	82.35%	82.35%	82.35%	86.36%	13.41%	+
50% ± 4%	46%	54%	82.35%	82.35%	82.35%	86.67%	13.72%	+
50% ± 5%	45%	55%	82.35%	82.35%	82.35%	87.23%	14.33%	+
50% ± 6%	44%	56%	78.95%	78.95%	78.95%	84.62%	15.85%	+
50% ± 7%	43%	57%	78.95%	71.43%	75.00%	82.14%	17.07%	+
50% ± 8%	42%	58%	69.57%	69.57%	69.57%	78.12%	19.51%	+
50% ± 9%	41%	59%	69.57%	66.67%	68.09%	78.26%	21.04%	+
50% ± 10%	40%	60%	70.83%	60.71%	65.38%	76.00%	22.87%	+
50% ± 11%	39%	61%	68.00%	58.62%	62.96%	74.03%	23.48%	+
50% ± 12%	38%	62%	65.38%	56.67%	60.71%	72.84%	24.70%	+
50% ± 13%	37%	63%	65.52%	57.58%	61.29%	72.73%	26.83%	+
50% ± 14%	36%	64%	64.52%	58.82%	61.54%	72.53%	27.74%	+
50% ± 15%	35%	65%	66.67%	57.89%	61.97%	72.45%	29.88%	+
50% ± 16%	34%	66%	65.71%	54.76%	59.74%	70.75%	32.32%	+
50% ± 17%	33%	67%	63.89%	54.76%	58.97%	70.37%	32.93%	+
50% ± 18%	32%	68%	63.16%	54.55%	58.54%	69.37%	33.84%	+
50% ± 19%	31%	69%	64.10%	52.08%	57.47%	68.64%	35.98%	+
50% ± 20%	30%	70%	66.67%	51.85%	58.33%	68.25%	38.41%	+
50% ± 21%	29%	71%	65.91%	52.73%	58.59%	68.70%	39.94%	+
50% ± 22%	28%	72%	62.00%	54.39%	57.94%	67.39%	42.07%	+
50% ± 23%	27%	73%	62.75%	55.17%	58.72%	68.75%	43.90%	+
50% ± 24%	26%	74%	61.82%	55.74%	58.62%	68.21%	46.04%	+
50% ± 25%	25%	75%	60.34%	54.69%	57.38%	67.09%	48.17%	+
50% ± 26%	24%	76%	59.38%	55.88%	57.58%	66.47%	50.91%	+
50% ± 27%	23%	77%	57.35%	55.71%	56.52%	65.32%	52.74%	+
50% ± 28%	22%	78%	56.52%	54.93%	55.71%	64.57%	53.35%	+
50% ± 29%	21%	79%	56.52%	54.93%	55.71%	64.57%	53.35%	+
50% ± 30%	20%	80%	56.52%	54.17%	55.32%	64.61%	54.27%	+
50% ± 31%	19%	81%	54.17%	54.17%	54.17%	63.93%	55.79%	+
50% ± 32%	18%	82%	52.63%	54.05%	53.33%	62.96%	57.62%	+
50% ± 33%	17%	83%	52.56%	53.95%	53.25%	62.50%	58.54%	+
50% ± 34%	16%	84%	52.56%	53.95%	53.25%	62.69%	58.84%	+
50% ± 35%	15%	85%	51.90%	53.95%	52.90%	62.76%	59.76%	+
50% ± 36%	14%	86%	53.09%	55.13%	54.09%	63.32%	60.67%	+
50% ± 37%	13%	87%	52.44%	55.13%	53.75%	63.37%	61.59%	+
50% ± 38%	12%	88%	52.38%	55.00%	53.66%	63.29%	63.11%	+
50% ± 39%	11%	89%	52.38%	55.00%	53.66%	64.49%	65.24%	+
50% ± 40%	10%	90%	52.94%	55.56%	54.22%	64.81%	65.85%	+
50% ± 41%	9%	91%	53.49%	55.42%	54.44%	64.84%	66.77%	+
50% ± 42%	8%	92%	53.93%	55.81%	54.86%	64.57%	67.99%	+
50% ± 43%	7%	93%	52.75%	55.17%	53.93%	63.88%	69.21%	+
50% ± 44%	6%	94%	52.63%	56.18%	54.35%	63.64%	70.43%	+
50% ± 45%	5%	95%	51.00%	56.67%	53.68%	62.71%	71.95%	+
50% ± 46%	4%	96%	52.83%	58.95%	55.72%	63.37%	74.09%	+
50% ± 47%	3%	97%	50.45%	58.33%	54.11%	61.85%	75.91%	+
50% ± 48%	2%	98%	50.00%	58.33%	53.85%	61.60%	76.22%	+
50% ± 49%	1%	99%	49.14%	58.76%	53.52%	61.33%	78.05%	+
50% ± 50%	0%	100%	50.00%	57.52%	53.50%	60.35%	86.89%	+

Table B.3 How the performance changes when varying the PPC range (the 3^{rd} run).

Range	Lower	Upper	Precision	Recall	F1	Accuracy	Coverage	Bayes Factor
50% ± 1%	49%	51%	83.33%	100.00%	90.91%	94.29%	10.67%	+
50% ± 2%	48%	52%	80.00%	85.71%	82.76%	87.50%	12.20%	+
50% ± 3%	47%	53%	82.35%	82.35%	82.35%	86.36%	13.41%	+
50% ± 4%	46%	54%	82.35%	82.35%	82.35%	86.67%	13.72%	+
50% ± 5%	45%	55%	82.35%	82.35%	82.35%	87.23%	14.33%	+
50% ± 6%	44%	56%	78.95%	78.95%	78.95%	84.62%	15.85%	+
50% ± 7%	43%	57%	78.95%	71.43%	75.00%	82.14%	17.07%	+
50% ± 8%	42%	58%	69.57%	69.57%	69.57%	78.12%	19.51%	+
50% ± 9%	41%	59%	69.57%	66.67%	68.09%	78.26%	21.04%	+
50% ± 10%	40%	60%	70.83%	60.71%	65.38%	76.00%	22.87%	+
50% ± 11%	39%	61%	68.00%	58.62%	62.96%	74.03%	23.48%	+
50% ± 12%	38%	62%	65.38%	56.67%	60.71%	72.84%	24.70%	+
50% ± 13%	37%	63%	65.52%	57.58%	61.29%	72.73%	26.83%	+
50% ± 14%	36%	64%	64.52%	58.82%	61.54%	72.53%	27.74%	+
50% ± 15%	35%	65%	66.67%	57.89%	61.97%	72.45%	29.88%	+
50% ± 16%	34%	66%	65.71%	54.76%	59.74%	70.75%	32.32%	+
50% ± 17%	33%	67%	63.89%	54.76%	58.97%	70.37%	32.93%	+
50% ± 18%	32%	68%	63.16%	54.55%	58.54%	69.37%	33.84%	+
50% ± 19%	31%	69%	64.10%	52.08%	57.47%	68.64%	35.98%	+
50% ± 20%	30%	70%	66.67%	51.85%	58.33%	68.25%	38.41%	+
50% ± 21%	29%	71%	65.91%	52.73%	58.59%	68.70%	39.94%	+
50% ± 22%	28%	72%	62.00%	54.39%	57.94%	67.39%	42.07%	+
50% ± 23%	27%	73%	62.75%	55.17%	58.72%	68.75%	43.90%	+
50% ± 24%	26%	74%	61.82%	55.74%	58.62%	68.21%	46.04%	+
50% ± 25%	25%	75%	60.34%	54.69%	57.38%	67.09%	48.17%	+
50% ± 26%	24%	76%	59.38%	55.88%	57.58%	66.47%	50.91%	+
50% ± 27%	23%	77%	57.35%	55.71%	56.52%	65.32%	52.74%	+
50% ± 28%	22%	78%	56.52%	54.93%	55.71%	64.57%	53.35%	+
50% ± 29%	21%	79%	56.52%	54.93%	55.71%	64.57%	53.35%	+
50% ± 30%	20%	80%	56.52%	54.17%	55.32%	64.61%	54.27%	+
50% ± 31%	19%	81%	54.17%	54.17%	54.17%	63.93%	55.79%	+
50% ± 32%	18%	82%	52.63%	54.05%	53.33%	62.96%	57.62%	+
50% ± 33%	17%	83%	52.56%	53.95%	53.25%	62.50%	58.54%	+
50% ± 34%	16%	84%	52.56%	53.95%	53.25%	62.69%	58.84%	+
50% ± 35%	15%	85%	51.90%	53.95%	52.90%	62.76%	59.76%	+
50% ± 36%	14%	86%	53.09%	55.13%	54.09%	63.32%	60.67%	+
50% ± 37%	13%	87%	52.44%	55.13%	53.75%	63.37%	61.59%	+
50% ± 38%	12%	88%	52.38%	55.00%	53.66%	63.29%	63.11%	+
50% ± 39%	11%	89%	52.38%	55.00%	53.66%	64.49%	65.24%	+
50% ± 40%	10%	90%	52.94%	55.56%	54.22%	64.81%	65.85%	+
50% ± 41%	9%	91%	53.49%	55.42%	54.44%	64.84%	66.77%	+
50% ± 42%	8%	92%	53.93%	55.81%	54.86%	64.57%	67.99%	+
50% ± 43%	7%	93%	52.75%	55.17%	53.93%	63.88%	69.21%	+
50% ± 44%	6%	94%	52.63%	56.18%	54.35%	63.64%	70.43%	+
50% ± 45%	5%	95%	51.00%	56.67%	53.68%	62.71%	71.95%	+
50% ± 46%	4%	96%	52.83%	58.95%	55.72%	63.37%	74.09%	+
50% ± 47%	3%	97%	50.45%	58.33%	54.11%	61.85%	75.91%	+
50% ± 48%	2%	98%	50.00%	58.33%	53.85%	61.60%	76.22%	+
50% ± 49%	1%	99%	49.14%	58.76%	53.52%	61.33%	78.05%	+
50% ± 50%	0%	100%	50.00%	57.52%	53.50%	60.35%	86.89%	+

Table B.4 How the performance changes when varying the PPC range (the 4^{th} run).

Range	**Lower**	**Upper**	**Precision**	**Recall**	**F1**	**Accuracy**	**Coverage**	**Bayes Factor**
50% ± 1%	49%	51%	83.33%	100.00%	90.91%	94.29%	10.67%	+
50% ± 2%	48%	52%	80.00%	85.71%	82.76%	87.50%	12.20%	+
50% ± 3%	47%	53%	82.35%	82.35%	82.35%	86.36%	13.41%	+
50% ± 4%	46%	54%	82.35%	82.35%	82.35%	86.67%	13.72%	+
50% ± 5%	45%	55%	82.35%	82.35%	82.35%	87.23%	14.33%	+
50% ± 6%	44%	56%	78.95%	78.95%	78.95%	84.62%	15.85%	+
50% ± 7%	43%	57%	78.95%	71.43%	75.00%	82.14%	17.07%	+
50% ± 8%	42%	58%	69.57%	69.57%	69.57%	78.12%	19.51%	+
50% ± 9%	41%	59%	69.57%	66.67%	68.09%	78.26%	21.04%	+
50% ± 10%	40%	60%	70.83%	60.71%	65.38%	76.00%	22.87%	+
50% ± 11%	39%	61%	68.00%	58.62%	62.96%	74.03%	23.48%	+
50% ± 12%	38%	62%	65.38%	56.67%	60.71%	72.84%	24.70%	+
50% ± 13%	37%	63%	65.52%	57.58%	61.29%	72.73%	26.83%	+
50% ± 14%	36%	64%	64.52%	58.82%	61.54%	72.53%	27.74%	+
50% ± 15%	35%	65%	66.67%	57.89%	61.97%	72.45%	29.88%	+
50% ± 16%	34%	66%	65.71%	54.76%	59.74%	70.75%	32.32%	+
50% ± 17%	33%	67%	63.89%	54.76%	58.97%	70.37%	32.93%	+
50% ± 18%	32%	68%	63.16%	54.55%	58.54%	69.37%	33.84%	+
50% ± 19%	31%	69%	64.10%	52.08%	57.47%	68.64%	35.98%	+
50% ± 20%	30%	70%	66.67%	51.85%	58.33%	68.25%	38.41%	+
50% ± 21%	29%	71%	65.91%	52.73%	58.59%	68.70%	39.94%	+
50% ± 22%	28%	72%	62.00%	54.39%	57.94%	67.39%	42.07%	+
50% ± 23%	27%	73%	62.75%	55.17%	58.72%	68.75%	43.90%	+
50% ± 24%	26%	74%	61.82%	55.74%	58.62%	68.21%	46.04%	+
50% ± 25%	25%	75%	60.34%	54.69%	57.38%	67.09%	48.17%	+
50% ± 26%	24%	76%	59.38%	55.88%	57.58%	66.47%	50.91%	+
50% ± 27%	23%	77%	57.35%	55.71%	56.52%	65.32%	52.74%	+
50% ± 28%	22%	78%	56.52%	54.93%	55.71%	64.57%	53.35%	+
50% ± 29%	21%	79%	56.52%	54.93%	55.71%	64.57%	53.35%	+
50% ± 30%	20%	80%	56.52%	54.17%	55.32%	64.61%	54.27%	+
50% ± 31%	19%	81%	54.17%	54.17%	54.17%	63.93%	55.79%	+
50% ± 32%	18%	82%	52.63%	54.05%	53.33%	62.96%	57.62%	+
50% ± 33%	17%	83%	52.56%	53.95%	53.25%	62.50%	58.54%	+
50% ± 34%	16%	84%	52.56%	53.95%	53.25%	62.69%	58.84%	+
50% ± 35%	15%	85%	51.90%	53.95%	52.90%	62.76%	59.76%	+
50% ± 36%	14%	86%	53.09%	55.13%	54.09%	63.32%	60.67%	+
50% ± 37%	13%	87%	52.44%	55.13%	53.75%	63.37%	61.59%	+
50% ± 38%	12%	88%	52.38%	55.00%	53.66%	63.29%	63.11%	+
50% ± 39%	11%	89%	52.38%	55.00%	53.66%	64.49%	65.24%	+
50% ± 40%	10%	90%	52.94%	55.56%	54.22%	64.81%	65.85%	+
50% ± 41%	9%	91%	53.49%	55.42%	54.44%	64.84%	66.77%	+
50% ± 42%	8%	92%	53.93%	55.81%	54.86%	64.57%	67.99%	+
50% ± 43%	7%	93%	52.75%	55.17%	53.93%	63.88%	69.21%	+
50% ± 44%	6%	94%	52.63%	56.18%	54.35%	63.64%	70.43%	+
50% ± 45%	5%	95%	51.00%	56.67%	53.68%	62.71%	71.95%	+
50% ± 46%	4%	96%	52.83%	58.95%	55.72%	63.37%	74.09%	+
50% ± 47%	3%	97%	50.45%	58.33%	54.11%	61.85%	75.91%	+
50% ± 48%	2%	98%	50.00%	58.33%	53.85%	61.60%	76.22%	+
50% ± 49%	1%	99%	49.14%	58.76%	53.52%	61.33%	78.05%	+
50% ± 50%	0%	100%	50.00%	57.52%	53.50%	60.35%	86.89%	+

Table B.5 How the performance changes when varying the PPC range (the 5^{th} run).

Range	Lower	Upper	Precision	Recall	F1	Accuracy	Coverage	Bayes Factor
50% ± 1%	49%	51%	83.33%	100.00%	90.91%	94.29%	10.67%	+
50% ± 2%	48%	52%	80.00%	85.71%	82.76%	87.50%	12.20%	+
50% ± 3%	47%	53%	82.35%	82.35%	82.35%	86.36%	13.41%	+
50% ± 4%	46%	54%	82.35%	82.35%	82.35%	86.67%	13.72%	+
50% ± 5%	45%	55%	82.35%	82.35%	82.35%	87.23%	14.33%	+
50% ± 6%	44%	56%	78.95%	78.95%	78.95%	84.62%	15.85%	+
50% ± 7%	43%	57%	78.95%	71.43%	75.00%	82.14%	17.07%	+
50% ± 8%	42%	58%	69.57%	69.57%	69.57%	78.12%	19.51%	+
50% ± 9%	41%	59%	69.57%	66.67%	68.09%	78.26%	21.04%	+
50% ± 10%	40%	60%	70.83%	60.71%	65.38%	76.00%	22.87%	+
50% ± 11%	39%	61%	68.00%	58.62%	62.96%	74.03%	23.48%	+
50% ± 12%	38%	62%	65.38%	56.67%	60.71%	72.84%	24.70%	+
50% ± 13%	37%	63%	65.52%	57.58%	61.29%	72.73%	26.83%	+
50% ± 14%	36%	64%	64.52%	58.82%	61.54%	72.53%	27.74%	+
50% ± 15%	35%	65%	66.67%	57.89%	61.97%	72.45%	29.88%	+
50% ± 16%	34%	66%	65.71%	54.76%	59.74%	70.75%	32.32%	+
50% ± 17%	33%	67%	63.89%	54.76%	58.97%	70.37%	32.93%	+
50% ± 18%	32%	68%	63.16%	54.55%	58.54%	69.37%	33.84%	+
50% ± 19%	31%	69%	64.10%	52.08%	57.47%	68.64%	35.98%	+
50% ± 20%	30%	70%	66.67%	51.85%	58.33%	68.25%	38.41%	+
50% ± 21%	29%	71%	65.91%	52.73%	58.59%	68.70%	39.94%	+
50% ± 22%	28%	72%	62.00%	54.39%	57.94%	67.39%	42.07%	+
50% ± 23%	27%	73%	62.75%	55.17%	58.72%	68.75%	43.90%	+
50% ± 24%	26%	74%	61.82%	55.74%	58.62%	68.21%	46.04%	+
50% ± 25%	25%	75%	60.34%	54.69%	57.38%	67.09%	48.17%	+
50% ± 26%	24%	76%	59.38%	55.88%	57.58%	66.47%	50.91%	+
50% ± 27%	23%	77%	57.35%	55.71%	56.52%	65.32%	52.74%	+
50% ± 28%	22%	78%	56.52%	54.93%	55.71%	64.57%	53.35%	+
50% ± 29%	21%	79%	56.52%	54.93%	55.71%	64.57%	53.35%	+
50% ± 30%	20%	80%	56.52%	54.17%	55.32%	64.61%	54.27%	+
50% ± 31%	19%	81%	54.17%	54.17%	54.17%	63.93%	55.79%	+
50% ± 32%	18%	82%	52.63%	54.05%	53.33%	62.96%	57.62%	+
50% ± 33%	17%	83%	52.56%	53.95%	53.25%	62.50%	58.54%	+
50% ± 34%	16%	84%	52.56%	53.95%	53.25%	62.69%	58.84%	+
50% ± 35%	15%	85%	51.90%	53.95%	52.90%	62.76%	59.76%	+
50% ± 36%	14%	86%	53.09%	55.13%	54.09%	63.32%	60.67%	+
50% ± 37%	13%	87%	52.44%	55.13%	53.75%	63.37%	61.59%	+
50% ± 38%	12%	88%	52.38%	55.00%	53.66%	63.29%	63.11%	+
50% ± 39%	11%	89%	52.38%	55.00%	53.66%	64.49%	65.24%	+
50% ± 40%	10%	90%	52.94%	55.56%	54.22%	64.81%	65.85%	+
50% ± 41%	9%	91%	53.49%	55.42%	54.44%	64.84%	66.77%	+
50% ± 42%	8%	92%	53.93%	55.81%	54.86%	64.57%	67.99%	+
50% ± 43%	7%	93%	52.75%	55.17%	53.93%	63.88%	69.21%	+
50% ± 44%	6%	94%	52.63%	56.18%	54.35%	63.64%	70.43%	+
50% ± 45%	5%	95%	51.00%	56.67%	53.68%	62.71%	71.95%	+
50% ± 46%	4%	96%	52.83%	58.95%	55.72%	63.37%	74.09%	+
50% ± 47%	3%	97%	50.45%	58.33%	54.11%	61.85%	75.91%	+
50% ± 48%	2%	98%	50.00%	58.33%	53.85%	61.60%	76.22%	+
50% ± 49%	1%	99%	49.14%	58.76%	53.52%	61.33%	78.05%	+
50% ± 50%	0%	100%	50.00%	57.52%	53.50%	60.35%	86.89%	+

Table B.6 How the performance changes when varying the PPC range (the 6^{th} run).

Range	Lower	Upper	Precision	Recall	F1	Accuracy	Coverage	Bayes Factor
50% ± 1%	49%	51%	83.33%	100.00%	90.91%	94.29%	10.67%	+
50% ± 2%	48%	52%	80.00%	85.71%	82.76%	87.50%	12.20%	+
50% ± 3%	47%	53%	82.35%	82.35%	82.35%	86.36%	13.41%	+
50% ± 4%	46%	54%	82.35%	82.35%	82.35%	86.67%	13.72%	+
50% ± 5%	45%	55%	82.35%	82.35%	82.35%	87.23%	14.33%	+
50% ± 6%	44%	56%	78.95%	78.95%	78.95%	84.62%	15.85%	+
50% ± 7%	43%	57%	78.95%	71.43%	75.00%	82.14%	17.07%	+
50% ± 8%	42%	58%	69.57%	69.57%	69.57%	78.12%	19.51%	+
50% ± 9%	41%	59%	69.57%	66.67%	68.09%	78.26%	21.04%	+
50% ± 10%	40%	60%	70.83%	60.71%	65.38%	76.00%	22.87%	+
50% ± 11%	39%	61%	68.00%	58.62%	62.96%	74.03%	23.48%	+
50% ± 12%	38%	62%	65.38%	56.67%	60.71%	72.84%	24.70%	+
50% ± 13%	37%	63%	65.52%	57.58%	61.29%	72.73%	26.83%	+
50% ± 14%	36%	64%	64.52%	58.82%	61.54%	72.53%	27.74%	+
50% ± 15%	35%	65%	66.67%	57.89%	61.97%	72.45%	29.88%	+
50% ± 16%	34%	66%	65.71%	54.76%	59.74%	70.75%	32.32%	+
50% ± 17%	33%	67%	63.89%	54.76%	58.97%	70.37%	32.93%	+
50% ± 18%	32%	68%	63.16%	54.55%	58.54%	69.37%	33.84%	+
50% ± 19%	31%	69%	64.10%	52.08%	57.47%	68.64%	35.98%	+
50% ± 20%	30%	70%	66.67%	51.85%	58.33%	68.25%	38.41%	+
50% ± 21%	29%	71%	65.91%	52.73%	58.59%	68.70%	39.94%	+
50% ± 22%	28%	72%	62.00%	54.39%	57.94%	67.39%	42.07%	+
50% ± 23%	27%	73%	62.75%	55.17%	58.72%	68.75%	43.90%	+
50% ± 24%	26%	74%	61.82%	55.74%	58.62%	68.21%	46.04%	+
50% ± 25%	25%	75%	60.34%	54.69%	57.38%	67.09%	48.17%	+
50% ± 26%	24%	76%	59.38%	55.88%	57.58%	66.47%	50.91%	+
50% ± 27%	23%	77%	57.35%	55.71%	56.52%	65.32%	52.74%	+
50% ± 28%	22%	78%	56.52%	54.93%	55.71%	64.57%	53.35%	+
50% ± 29%	21%	79%	56.52%	54.93%	55.71%	64.57%	53.35%	+
50% ± 30%	20%	80%	56.52%	54.17%	55.32%	64.61%	54.27%	+
50% ± 31%	19%	81%	54.17%	54.17%	54.17%	63.93%	55.79%	+
50% ± 32%	18%	82%	52.63%	54.05%	53.33%	62.96%	57.62%	+
50% ± 33%	17%	83%	52.56%	53.95%	53.25%	62.50%	58.54%	+
50% ± 34%	16%	84%	52.56%	53.95%	53.25%	62.69%	58.84%	+
50% ± 35%	15%	85%	51.90%	53.95%	52.90%	62.76%	59.76%	+
50% ± 36%	14%	86%	53.09%	55.13%	54.09%	63.32%	60.67%	+
50% ± 37%	13%	87%	52.44%	55.13%	53.75%	63.37%	61.59%	+
50% ± 38%	12%	88%	52.38%	55.00%	53.66%	63.29%	63.11%	+
50% ± 39%	11%	89%	52.38%	55.00%	53.66%	64.49%	65.24%	+
50% ± 40%	10%	90%	52.94%	55.56%	54.22%	64.81%	65.85%	+
50% ± 41%	9%	91%	53.49%	55.42%	54.44%	64.84%	66.77%	+
50% ± 42%	8%	92%	53.93%	55.81%	54.86%	64.57%	67.99%	+
50% ± 43%	7%	93%	52.75%	55.17%	53.93%	63.88%	69.21%	+
50% ± 44%	6%	94%	52.63%	56.18%	54.35%	63.64%	70.43%	+
50% ± 45%	5%	95%	51.00%	56.67%	53.68%	62.71%	71.95%	+
50% ± 46%	4%	96%	52.83%	58.95%	55.72%	63.37%	74.09%	+
50% ± 47%	3%	97%	50.45%	58.33%	54.11%	61.85%	75.91%	+
50% ± 48%	2%	98%	50.00%	58.33%	53.85%	61.60%	76.22%	+
50% ± 49%	1%	99%	49.14%	58.76%	53.52%	61.33%	78.05%	+
50% ± 50%	0%	100%	50.00%	57.52%	53.50%	60.35%	86.89%	+

Table B.7 How the performance changes when varying the PPC range (the 7^{th} run).

Range	Lower	Upper	Precision	Recall	F1	Accuracy	Coverage	Bayes Factor
50% ± 1%	49%	51%	83.33%	100.00%	90.91%	94.29%	10.67%	+
50% ± 2%	48%	52%	80.00%	85.71%	82.76%	87.50%	12.20%	+
50% ± 3%	47%	53%	82.35%	82.35%	82.35%	86.36%	13.41%	+
50% ± 4%	46%	54%	82.35%	82.35%	82.35%	86.67%	13.72%	+
50% ± 5%	45%	55%	82.35%	82.35%	82.35%	87.23%	14.33%	+
50% ± 6%	44%	56%	78.95%	78.95%	78.95%	84.62%	15.85%	+
50% ± 7%	43%	57%	78.95%	71.43%	75.00%	82.14%	17.07%	+
50% ± 8%	42%	58%	69.57%	69.57%	69.57%	78.12%	19.51%	+
50% ± 9%	41%	59%	69.57%	66.67%	68.09%	78.26%	21.04%	+
50% ± 10%	40%	60%	70.83%	60.71%	65.38%	76.00%	22.87%	+
50% ± 11%	39%	61%	68.00%	58.62%	62.96%	74.03%	23.48%	+
50% ± 12%	38%	62%	65.38%	56.67%	60.71%	72.84%	24.70%	+
50% ± 13%	37%	63%	65.52%	57.58%	61.29%	72.73%	26.83%	+
50% ± 14%	36%	64%	64.52%	58.82%	61.54%	72.53%	27.74%	+
50% ± 15%	35%	65%	66.67%	57.89%	61.97%	72.45%	29.88%	+
50% ± 16%	34%	66%	65.71%	54.76%	59.74%	70.75%	32.32%	+
50% ± 17%	33%	67%	63.89%	54.76%	58.97%	70.37%	32.93%	+
50% ± 18%	32%	68%	63.16%	54.55%	58.54%	69.37%	33.84%	+
50% ± 19%	31%	69%	64.10%	52.08%	57.47%	68.64%	35.98%	+
50% ± 20%	30%	70%	66.67%	51.85%	58.33%	68.25%	38.41%	+
50% ± 21%	29%	71%	65.91%	52.73%	58.59%	68.70%	39.94%	+
50% ± 22%	28%	72%	62.00%	54.39%	57.94%	67.39%	42.07%	+
50% ± 23%	27%	73%	62.75%	55.17%	58.72%	68.75%	43.90%	+
50% ± 24%	26%	74%	61.82%	55.74%	58.62%	68.21%	46.04%	+
50% ± 25%	25%	75%	60.34%	54.69%	57.38%	67.09%	48.17%	+
50% ± 26%	24%	76%	59.38%	55.88%	57.58%	66.47%	50.91%	+
50% ± 27%	23%	77%	57.35%	55.71%	56.52%	65.32%	52.74%	+
50% ± 28%	22%	78%	56.52%	54.93%	55.71%	64.57%	53.35%	+
50% ± 29%	21%	79%	56.52%	54.93%	55.71%	64.57%	53.35%	+
50% ± 30%	20%	80%	56.52%	54.17%	55.32%	64.61%	54.27%	+
50% ± 31%	19%	81%	54.17%	54.17%	54.17%	63.93%	55.79%	+
50% ± 32%	18%	82%	52.63%	54.05%	53.33%	62.96%	57.62%	+
50% ± 33%	17%	83%	52.56%	53.95%	53.25%	62.50%	58.54%	+
50% ± 34%	16%	84%	52.56%	53.95%	53.25%	62.69%	58.84%	+
50% ± 35%	15%	85%	51.90%	53.95%	52.90%	62.76%	59.76%	+
50% ± 36%	14%	86%	53.09%	55.13%	54.09%	63.32%	60.67%	+
50% ± 37%	13%	87%	52.44%	55.13%	53.75%	63.37%	61.59%	+
50% ± 38%	12%	88%	52.38%	55.00%	53.66%	63.29%	63.11%	+
50% ± 39%	11%	89%	52.38%	55.00%	53.66%	64.49%	65.24%	+
50% ± 40%	10%	90%	52.94%	55.56%	54.22%	64.81%	65.85%	+
50% ± 41%	9%	91%	53.49%	55.42%	54.44%	64.84%	66.77%	+
50% ± 42%	8%	92%	53.93%	55.81%	54.86%	64.57%	67.99%	+
50% ± 43%	7%	93%	52.75%	55.17%	53.93%	63.88%	69.21%	+
50% ± 44%	6%	94%	52.63%	56.18%	54.35%	63.64%	70.43%	+
50% ± 45%	5%	95%	51.00%	56.67%	53.68%	62.71%	71.95%	+
50% ± 46%	4%	96%	52.83%	58.95%	55.72%	63.37%	74.09%	+
50% ± 47%	3%	97%	50.45%	58.33%	54.11%	61.85%	75.91%	+
50% ± 48%	2%	98%	50.00%	58.33%	53.85%	61.60%	76.22%	+
50% ± 49%	1%	99%	49.14%	58.76%	53.52%	61.33%	78.05%	+
50% ± 50%	0%	100%	50.00%	57.52%	53.50%	60.35%	86.89%	+

Table B.8 How the performance changes when varying the PPC range (the 8^{th} run).

Range	Lower	Upper	Precision	Recall	F1	Accuracy	Coverage	Bayes Factor
50% ± 1%	49%	51%	83.33%	100.00%	90.91%	94.29%	10.67%	+
50% ± 2%	48%	52%	80.00%	85.71%	82.76%	87.50%	12.20%	+
50% ± 3%	47%	53%	82.35%	82.35%	82.35%	86.36%	13.41%	+
50% ± 4%	46%	54%	82.35%	82.35%	82.35%	86.67%	13.72%	+
50% ± 5%	45%	55%	82.35%	82.35%	82.35%	87.23%	14.33%	+
50% ± 6%	44%	56%	78.95%	78.95%	78.95%	84.62%	15.85%	+
50% ± 7%	43%	57%	78.95%	71.43%	75.00%	82.14%	17.07%	+
50% ± 8%	42%	58%	69.57%	69.57%	69.57%	78.12%	19.51%	+
50% ± 9%	41%	59%	69.57%	66.67%	68.09%	78.26%	21.04%	+
50% ± 10%	40%	60%	70.83%	60.71%	65.38%	76.00%	22.87%	+
50% ± 11%	39%	61%	68.00%	58.62%	62.96%	74.03%	23.48%	+
50% ± 12%	38%	62%	65.38%	56.67%	60.71%	72.84%	24.70%	+
50% ± 13%	37%	63%	65.52%	57.58%	61.29%	72.73%	26.83%	+
50% ± 14%	36%	64%	64.52%	58.82%	61.54%	72.53%	27.74%	+
50% ± 15%	35%	65%	66.67%	57.89%	61.97%	72.45%	29.88%	+
50% ± 16%	34%	66%	65.71%	54.76%	59.74%	70.75%	32.32%	+
50% ± 17%	33%	67%	63.89%	54.76%	58.97%	70.37%	32.93%	+
50% ± 18%	32%	68%	63.16%	54.55%	58.54%	69.37%	33.84%	+
50% ± 19%	31%	69%	64.10%	52.08%	57.47%	68.64%	35.98%	+
50% ± 20%	30%	70%	66.67%	51.85%	58.33%	68.25%	38.41%	+
50% ± 21%	29%	71%	65.91%	52.73%	58.59%	68.70%	39.94%	+
50% ± 22%	28%	72%	62.00%	54.39%	57.94%	67.39%	42.07%	+
50% ± 23%	27%	73%	62.75%	55.17%	58.72%	68.75%	43.90%	+
50% ± 24%	26%	74%	61.82%	55.74%	58.62%	68.21%	46.04%	+
50% ± 25%	25%	75%	60.34%	54.69%	57.38%	67.09%	48.17%	+
50% ± 26%	24%	76%	59.38%	55.88%	57.58%	66.47%	50.91%	+
50% ± 27%	23%	77%	57.35%	55.71%	56.52%	65.32%	52.74%	+
50% ± 28%	22%	78%	56.52%	54.93%	55.71%	64.57%	53.35%	+
50% ± 29%	21%	79%	56.52%	54.93%	55.71%	64.57%	53.35%	+
50% ± 30%	20%	80%	56.52%	54.17%	55.32%	64.61%	54.27%	+
50% ± 31%	19%	81%	54.17%	54.17%	54.17%	63.93%	55.79%	+
50% ± 32%	18%	82%	52.63%	54.05%	53.33%	62.96%	57.62%	+
50% ± 33%	17%	83%	52.56%	53.95%	53.25%	62.50%	58.54%	+
50% ± 34%	16%	84%	52.56%	53.95%	53.25%	62.69%	58.84%	+
50% ± 35%	15%	85%	51.90%	53.95%	52.90%	62.76%	59.76%	+
50% ± 36%	14%	86%	53.09%	55.13%	54.09%	63.32%	60.67%	+
50% ± 37%	13%	87%	52.44%	55.13%	53.75%	63.37%	61.59%	+
50% ± 38%	12%	88%	52.38%	55.00%	53.66%	63.29%	63.11%	+
50% ± 39%	11%	89%	52.38%	55.00%	53.66%	64.49%	65.24%	+
50% ± 40%	10%	90%	52.94%	55.56%	54.22%	64.81%	65.85%	+
50% ± 41%	9%	91%	53.49%	55.42%	54.44%	64.84%	66.77%	+
50% ± 42%	8%	92%	53.93%	55.81%	54.86%	64.57%	67.99%	+
50% ± 43%	7%	93%	52.75%	55.17%	53.93%	63.88%	69.21%	+
50% ± 44%	6%	94%	52.63%	56.18%	54.35%	63.64%	70.43%	+
50% ± 45%	5%	95%	51.00%	56.67%	53.68%	62.71%	71.95%	+
50% ± 46%	4%	96%	52.83%	58.95%	55.72%	63.37%	74.09%	+
50% ± 47%	3%	97%	50.45%	58.33%	54.11%	61.85%	75.91%	+
50% ± 48%	2%	98%	50.00%	58.33%	53.85%	61.60%	76.22%	+
50% ± 49%	1%	99%	49.14%	58.76%	53.52%	61.33%	78.05%	+
50% ± 50%	0%	100%	50.00%	57.52%	53.50%	60.35%	86.89%	+

Table B.9 How the performance changes when varying the PPC range (the 9^{th} run).

Range	Lower	Upper	Precision	Recall	F1	Accuracy	Coverage	Bayes Factor
50% ± 1%	49%	51%	83.33%	100.00%	90.91%	94.29%	10.67%	+
50% ± 2%	48%	52%	80.00%	85.71%	82.76%	87.50%	12.20%	+
50% ± 3%	47%	53%	82.35%	82.35%	82.35%	86.36%	13.41%	+
50% ± 4%	46%	54%	82.35%	82.35%	82.35%	86.67%	13.72%	+
50% ± 5%	45%	55%	82.35%	82.35%	82.35%	87.23%	14.33%	+
50% ± 6%	44%	56%	78.95%	78.95%	78.95%	84.62%	15.85%	+
50% ± 7%	43%	57%	78.95%	71.43%	75.00%	82.14%	17.07%	+
50% ± 8%	42%	58%	69.57%	69.57%	69.57%	78.12%	19.51%	+
50% ± 9%	41%	59%	69.57%	66.67%	68.09%	78.26%	21.04%	+
50% ± 10%	40%	60%	70.83%	60.71%	65.38%	76.00%	22.87%	+
50% ± 11%	39%	61%	68.00%	58.62%	62.96%	74.03%	23.48%	+
50% ± 12%	38%	62%	65.38%	56.67%	60.71%	72.84%	24.70%	+
50% ± 13%	37%	63%	65.52%	57.58%	61.29%	72.73%	26.83%	+
50% ± 14%	36%	64%	64.52%	58.82%	61.54%	72.53%	27.74%	+
50% ± 15%	35%	65%	66.67%	57.89%	61.97%	72.45%	29.88%	+
50% ± 16%	34%	66%	65.71%	54.76%	59.74%	70.75%	32.32%	+
50% ± 17%	33%	67%	63.89%	54.76%	58.97%	70.37%	32.93%	+
50% ± 18%	32%	68%	63.16%	54.55%	58.54%	69.37%	33.84%	+
50% ± 19%	31%	69%	64.10%	52.08%	57.47%	68.64%	35.98%	+
50% ± 20%	30%	70%	66.67%	51.85%	58.33%	68.25%	38.41%	+
50% ± 21%	29%	71%	65.91%	52.73%	58.59%	68.70%	39.94%	+
50% ± 22%	28%	72%	62.00%	54.39%	57.94%	67.39%	42.07%	+
50% ± 23%	27%	73%	62.75%	55.17%	58.72%	68.75%	43.90%	+
50% ± 24%	26%	74%	61.82%	55.74%	58.62%	68.21%	46.04%	+
50% ± 25%	25%	75%	60.34%	54.69%	57.38%	67.09%	48.17%	+
50% ± 26%	24%	76%	59.38%	55.88%	57.58%	66.47%	50.91%	+
50% ± 27%	23%	77%	57.35%	55.71%	56.52%	65.32%	52.74%	+
50% ± 28%	22%	78%	56.52%	54.93%	55.71%	64.57%	53.35%	+
50% ± 29%	21%	79%	56.52%	54.93%	55.71%	64.57%	53.35%	+
50% ± 30%	20%	80%	56.52%	54.17%	55.32%	64.61%	54.27%	+
50% ± 31%	19%	81%	54.17%	54.17%	54.17%	63.93%	55.79%	+
50% ± 32%	18%	82%	52.63%	54.05%	53.33%	62.96%	57.62%	+
50% ± 33%	17%	83%	52.56%	53.95%	53.25%	62.50%	58.54%	+
50% ± 34%	16%	84%	52.56%	53.95%	53.25%	62.69%	58.84%	+
50% ± 35%	15%	85%	51.90%	53.95%	52.90%	62.76%	59.76%	+
50% ± 36%	14%	86%	53.09%	55.13%	54.09%	63.32%	60.67%	+
50% ± 37%	13%	87%	52.44%	55.13%	53.75%	63.37%	61.59%	+
50% ± 38%	12%	88%	52.38%	55.00%	53.66%	63.29%	63.11%	+
50% ± 39%	11%	89%	52.38%	55.00%	53.66%	64.49%	65.24%	+
50% ± 40%	10%	90%	52.94%	55.56%	54.22%	64.81%	65.85%	+
50% ± 41%	9%	91%	53.49%	55.42%	54.44%	64.84%	66.77%	+
50% ± 42%	8%	92%	53.93%	55.81%	54.86%	64.57%	67.99%	+
50% ± 43%	7%	93%	52.75%	55.17%	53.93%	63.88%	69.21%	+
50% ± 44%	6%	94%	52.63%	56.18%	54.35%	63.64%	70.43%	+
50% ± 45%	5%	95%	51.00%	56.67%	53.68%	62.71%	71.95%	+
50% ± 46%	4%	96%	52.83%	58.95%	55.72%	63.37%	74.09%	+
50% ± 47%	3%	97%	50.45%	58.33%	54.11%	61.85%	75.91%	+
50% ± 48%	2%	98%	50.00%	58.33%	53.85%	61.60%	76.22%	+
50% ± 49%	1%	99%	49.14%	58.76%	53.52%	61.33%	78.05%	+
50% ± 50%	0%	100%	50.00%	57.52%	53.50%	60.35%	86.89%	+

Table B.10 How the performance changes when varying the PPC range (the 10^{th} run).

Table B.11, B.12 and B.13 show the experimental results for 1(a) Data Augmentation — Bayesian Framework.

Run	Precision	Recall	F1	Accuracy	Coverage	Cost
1	50.00%	56.72%	53.15%	59.15%	100.00%	–
2	50.00%	56.72%	53.15%	59.15%	100.00%	–
3	50.00%	56.72%	53.15%	59.15%	100.00%	–
4	50.00%	56.72%	53.15%	59.15%	100.00%	–
5	50.00%	56.72%	53.15%	59.15%	100.00%	–
6	50.00%	56.72%	53.15%	59.15%	100.00%	–
7	50.00%	56.72%	53.15%	59.15%	100.00%	–
8	50.00%	56.72%	53.15%	59.15%	100.00%	–
9	50.00%	56.72%	53.15%	59.15%	100.00%	–
10	50.00%	56.72%	53.15%	59.15%	100.00%	–
Average	50.00%	56.72%	53.15%	59.15%	100.00%	–
SD	(0.00%)	(0.00%)	(0.00%)	(0.00%)	(0.00%)	–

Table B.11 Results of 1(a)(i) Data Augmentation — Bayesian (BF−, PPC−) are shown.

Run	Precision	Recall	F1	Accuracy	Coverage	Cost
1	83.33%	100.00%	90.91%	94.29%	10.67%	$2\lambda_e + 293\lambda_r$
2	83.33%	100.00%	90.91%	94.29%	10.67%	$2\lambda_e + 293\lambda_r$
3	83.33%	100.00%	90.91%	94.29%	10.67%	$2\lambda_e + 293\lambda_r$
4	83.33%	100.00%	90.91%	94.29%	10.67%	$2\lambda_e + 293\lambda_r$
5	83.33%	100.00%	90.91%	94.29%	10.67%	$2\lambda_e + 293\lambda_r$
6	83.33%	100.00%	90.91%	94.29%	10.67%	$2\lambda_e + 293\lambda_r$
7	83.33%	100.00%	90.91%	94.29%	10.67%	$2\lambda_e + 293\lambda_r$
8	83.33%	100.00%	90.91%	94.29%	10.67%	$2\lambda_e + 293\lambda_r$
9	83.33%	100.00%	90.91%	94.29%	10.67%	$2\lambda_e + 293\lambda_r$
10	83.33%	100.00%	90.91%	94.29%	10.67%	$2\lambda_e + 293\lambda_r$
Average	83.33%	100.00%	90.91%	94.29%	10.67%	$2\lambda_e + 293\lambda_r$
SD	(0.00%)	(0.00%)	(0.00%)	(0.00%)	(0.00%)	(0.00%)

Table B.12 Results of 1(a)(ii) Data Augmentation — Bayesian (BF+, PPC+) are shown.

Run	Precision	Recall	F1	Accuracy	Coverage	Cost
1	50.00%	57.52%	53.50%	60.35%	86.89%	$113\lambda_e + 43\lambda_r$
2	50.00%	57.52%	53.50%	60.35%	86.89%	$113\lambda_e + 43\lambda_r$
3	50.00%	57.52%	53.50%	60.35%	86.89%	$113\lambda_e + 43\lambda_r$
4	50.00%	57.52%	53.50%	60.35%	86.89%	$113\lambda_e + 43\lambda_r$
5	50.00%	57.52%	53.50%	60.35%	86.89%	$113\lambda_e + 43\lambda_r$
6	50.00%	57.52%	53.50%	60.35%	86.89%	$113\lambda_e + 43\lambda_r$
7	50.00%	57.52%	53.50%	60.35%	86.89%	$113\lambda_e + 43\lambda_r$
8	50.00%	57.52%	53.50%	60.35%	86.89%	$113\lambda_e + 43\lambda_r$
9	50.00%	57.52%	53.50%	60.35%	86.89%	$113\lambda_e + 43\lambda_r$
10	50.00%	57.52%	53.50%	60.35%	86.89%	$113\lambda_e + 43\lambda_r$
Average	50.00%	57.52%	53.50%	60.35%	86.89%	$113\lambda_e + 43\lambda_r$
SD	(0.00%)	(0.00%)	(0.00%)	(0.00%)	(0.00%)	(0.00%)

Table B.13 Results of 1(a)(ii) Data Augmentation — Bayesian (BF+, PPC−) are shown

B.2.2 1(b) Data Augmentation — BERT

Table B.14 shows the experimental results for 1(b) Data Augmentation — BERT.

Run	Precision	Recall	F1	Accuracy	Coverage	Cost
1	95.31%	91.04%	93.13%	94.51%	100.00%	–
2	93.98%	93.28%	93.63%	94.82%	100.00%	–
3	95.38%	92.54%	93.94%	95.12%	100.00%	–
4	94.03%	94.03%	94.03%	95.12%	100.00%	–
5	96.92%	94.03%	95.45%	96.34%	100.00%	–
6	94.03%	94.03%	94.03%	95.12%	100.00%	–
7	94.78%	94.78%	94.78%	95.73%	100.00%	–
8	96.03%	90.3%	93.08%	94.51%	100.00%	–
9	95.42%	93.28%	94.34%	95.43%	100.00%	–
10	94.74%	94.03%	94.38%	95.43%	100.00%	–
Average	95.06%	93.13%	94.08%	95.21%	100.00%	–
SD	(0.90%)	(1.37%)	(0.68%)	(0.53%)	(0.00%)	–

Table B.14 Results of 1(b) Data Augmentation — BERT are shown.

B.2.3 1(c) Data Augmentation — Bayesian Framework + BERT

Table B.15 and B.16 show the experimental results for 1(c) Bayesian+BERT.

Run	Precision	Recall	F1	Accuracy	Coverage	Cost
1	96.18%	94.03%	95.09%	96.04%	100.00%	–
2	98.39%	91.04%	94.57%	95.73%	100.00%	–
3	96.88%	92.54%	94.66%	95.73%	100.00%	–
4	96.18%	94.03%	95.09%	96.04%	100.00%	–
5	91.49%	96.27%	93.82%	94.82%	100.00%	–
6	93.98%	93.28%	93.63%	94.82%	100.00%	–
7	93.53%	97.01%	95.24%	96.04%	100.00%	–
8	94.66%	92.54%	93.58%	94.82%	100.00%	–
9	95.38%	92.54%	93.94%	95.12%	100.00%	–
10	92.70%	94.78%	93.73%	94.82%	100.00%	–
Average	94.94%	93.81%	94.34%	95.40%	100.00%	–
SD	(2.08%)	(1.83%)	(0.66%)	(0.56%)	(0.00%)	–

Table B.15 Results of 1(c) Data Augmentation — Bayesian+BERT (BF+, PPC+) are shown.

Run	Precision	Recall	F1	Accuracy	Coverage	Cost
1	64.33%	75.37%	69.42%	72.87%	100.00%	–
2	65.03%	79.10%	71.38%	74.09%	100.00%	–
3	75.84%	84.33%	79.86%	82.62%	100.00%	–
4	67.33%	75.37%	71.13%	75.00%	100.00%	–
5	69.54%	78.36%	73.68%	77.13%	100.00%	–
6	62.59%	68.66%	65.48%	70.43%	100.00%	–
7	67.08%	80.60%	73.22%	75.91%	100.00%	–
8	62.35%	75.37%	68.24%	71.34%	100.00%	–
9	60.78%	69.40%	64.81%	69.21%	100.00%	–
10	67.68%	82.84%	74.50%	76.83%	100.00%	–
Average	66.26%	76.94%	71.17%	74.54%	100.00%	–
SD	(4.35%)	(5.19%)	(4.49%)	(3.91%)	(0.00%)	–

Table B.16 Results of 1(c) Data Augmentation — Bayesian+BERT (BF+, PPC−) are shown.

B.2.4 2(b) Unsupervised Learning — Bayesian Framework

Range	Lower	Upper	Precision	Recall	F1	Accuracy	Coverage	Bayes Factor
50% ± 1%	49%	51%	0.00%	0.00%	0.00%	71.43%	2.13%	+
50% ± 2%	48%	52%	40.00%	50.00%	44.44%	58.33%	3.66%	+
50% ± 3%	47%	53%	57.14%	57.14%	57.14%	62.50%	4.88%	+
50% ± 4%	46%	54%	57.14%	50.00%	53.33%	63.16%	5.79%	+
50% ± 5%	45%	55%	57.14%	44.44%	50.00%	63.64%	6.71%	+
50% ± 6%	44%	56%	55.56%	45.45%	50.00%	64.29%	8.54%	+
50% ± 7%	43%	57%	55.56%	38.46%	45.45%	62.50%	9.76%	+
50% ± 8%	42%	58%	46.15%	40.00%	42.86%	60.98%	12.50%	+
50% ± 9%	41%	59%	46.15%	37.50%	41.38%	63.04%	14.02%	+
50% ± 10%	40%	60%	50.00%	35.00%	41.18%	61.54%	15.85%	+
50% ± 11%	39%	61%	46.67%	33.33%	38.89%	59.26%	16.46%	+
50% ± 12%	38%	62%	43.75%	31.82%	36.84%	59.32%	17.99%	+
50% ± 13%	37%	63%	47.37%	36.00%	40.91%	60.61%	20.12%	+
50% ± 14%	36%	64%	47.62%	38.46%	42.55%	60.87%	21.04%	+
50% ± 15%	35%	65%	52.17%	40.00%	45.28%	61.84%	23.17%	+
50% ± 16%	34%	66%	52.00%	38.24%	44.07%	60.71%	25.61%	+
50% ± 17%	33%	67%	48.15%	38.24%	42.62%	59.77%	26.52%	+
50% ± 18%	32%	68%	48.28%	37.84%	42.42%	58.24%	27.74%	+
50% ± 19%	31%	69%	50.00%	36.59%	42.25%	58.16%	29.88%	+
50% ± 20%	30%	70%	54.55%	38.30%	45.00%	58.49%	32.32%	+
50% ± 21%	29%	71%	54.29%	39.58%	45.78%	59.82%	34.15%	+
50% ± 22%	28%	72%	51.22%	42.00%	46.15%	59.17%	36.59%	+
50% ± 23%	27%	73%	52.38%	43.14%	47.31%	61.11%	38.41%	+
50% ± 24%	26%	74%	51.06%	43.64%	47.06%	60.00%	41.16%	+
50% ± 25%	25%	75%	50.00%	43.10%	46.30%	59.15%	43.29%	+
50% ± 26%	24%	76%	50.00%	45.16%	47.46%	58.94%	46.04%	+
50% ± 27%	23%	77%	49.18%	46.15%	47.62%	58.23%	48.17%	+
50% ± 28%	22%	78%	48.39%	45.45%	46.87%	57.50%	48.78%	+
50% ± 29%	21%	79%	48.39%	44.78%	46.51%	57.41%	49.39%	+
50% ± 30%	20%	80%	48.39%	44.12%	46.15%	57.58%	50.30%	+
50% ± 31%	19%	81%	46.15%	44.12%	45.11%	57.06%	51.83%	+
50% ± 32%	18%	82%	44.93%	44.29%	44.60%	56.25%	53.66%	+
50% ± 33%	17%	83%	45.07%	44.44%	44.76%	55.87%	54.57%	+
50% ± 34%	16%	84%	45.07%	44.44%	44.76%	56.11%	54.88%	+
50% ± 35%	15%	85%	44.44%	44.44%	44.44%	56.28%	55.79%	+
50% ± 36%	14%	86%	45.95%	45.95%	45.95%	56.99%	56.71%	+
50% ± 37%	13%	87%	45.33%	45.95%	45.64%	57.14%	57.62%	+
50% ± 38%	12%	88%	45.45%	46.05%	45.75%	57.22%	59.15%	+
50% ± 39%	11%	89%	44.87%	46.05%	45.45%	58.42%	61.59%	+
50% ± 40%	10%	90%	45.57%	46.75%	46.15%	58.82%	62.20%	+
50% ± 41%	9%	91%	46.25%	46.84%	46.54%	58.94%	63.11%	+
50% ± 42%	8%	92%	46.99%	47.56%	47.27%	58.77%	64.33%	+
50% ± 43%	7%	93%	46.51%	47.62%	47.06%	58.33%	65.85%	+
50% ± 44%	6%	94%	46.67%	48.84%	47.73%	58.18%	67.07%	+
50% ± 45%	5%	95%	44.79%	49.43%	46.99%	57.08%	68.90%	+
50% ± 46%	4%	96%	47.06%	52.17%	49.48%	57.94%	71.04%	+
50% ± 47%	3%	97%	44.86%	51.61%	48.00%	56.49%	72.87%	+
50% ± 48%	2%	98%	44.44%	51.61%	47.76%	56.25%	73.17%	+
50% ± 49%	1%	99%	43.36%	52.13%	47.34%	55.87%	75.30%	+
50% ± 50%	0%	100%	45.31%	51.33%	48.13%	55.36%	85.37%	+

Table B.17 How the performance changes when varying the PPC range (the 1^{st} run).

Range	Lower	Upper	Precision	Recall	F1	Accuracy	Coverage	Bayes Factor
50% ± 1%	49%	51%	0.00%	0.00%	0.00%	71.43%	2.13%	+
50% ± 2%	48%	52%	40.00%	50.00%	44.44%	58.33%	3.66%	+
50% ± 3%	47%	53%	57.14%	57.14%	57.14%	62.50%	4.88%	+
50% ± 4%	46%	54%	57.14%	50.00%	53.33%	63.16%	5.79%	+
50% ± 5%	45%	55%	57.14%	44.44%	50.00%	63.64%	6.71%	+
50% ± 6%	44%	56%	55.56%	45.45%	50.00%	64.29%	8.54%	+
50% ± 7%	43%	57%	55.56%	38.46%	45.45%	62.50%	9.76%	+
50% ± 8%	42%	58%	46.15%	40.00%	42.86%	60.98%	12.50%	+
50% ± 9%	41%	59%	46.15%	37.50%	41.38%	63.04%	14.02%	+
50% ± 10%	40%	60%	50.00%	35.00%	41.18%	61.54%	15.85%	+
50% ± 11%	39%	61%	46.67%	33.33%	38.89%	59.26%	16.46%	+
50% ± 12%	38%	62%	43.75%	31.82%	36.84%	59.32%	17.99%	+
50% ± 13%	37%	63%	47.37%	36.00%	40.91%	60.61%	20.12%	+
50% ± 14%	36%	64%	47.62%	38.46%	42.55%	60.87%	21.04%	+
50% ± 15%	35%	65%	52.17%	40.00%	45.28%	61.84%	23.17%	+
50% ± 16%	34%	66%	52.00%	38.24%	44.07%	60.71%	25.61%	+
50% ± 17%	33%	67%	48.15%	38.24%	42.62%	59.77%	26.52%	+
50% ± 18%	32%	68%	48.28%	37.84%	42.42%	58.24%	27.74%	+
50% ± 19%	31%	69%	50.00%	36.59%	42.25%	58.16%	29.88%	+
50% ± 20%	30%	70%	54.55%	38.30%	45.00%	58.49%	32.32%	+
50% ± 21%	29%	71%	54.29%	39.58%	45.78%	59.82%	34.15%	+
50% ± 22%	28%	72%	51.22%	42.00%	46.15%	59.17%	36.59%	+
50% ± 23%	27%	73%	52.38%	43.14%	47.31%	61.11%	38.41%	+
50% ± 24%	26%	74%	51.06%	43.64%	47.06%	60.00%	41.16%	+
50% ± 25%	25%	75%	50.00%	43.10%	46.30%	59.15%	43.29%	+
50% ± 26%	24%	76%	50.00%	45.16%	47.46%	58.94%	46.04%	+
50% ± 27%	23%	77%	49.18%	46.15%	47.62%	58.23%	48.17%	+
50% ± 28%	22%	78%	48.39%	45.45%	46.87%	57.50%	48.78%	+
50% ± 29%	21%	79%	48.39%	44.78%	46.51%	57.41%	49.39%	+
50% ± 30%	20%	80%	48.39%	44.12%	46.15%	57.58%	50.30%	+
50% ± 31%	19%	81%	46.15%	44.12%	45.11%	57.06%	51.83%	+
50% ± 32%	18%	82%	44.93%	44.29%	44.60%	56.25%	53.66%	+
50% ± 33%	17%	83%	45.07%	44.44%	44.76%	55.87%	54.57%	+
50% ± 34%	16%	84%	45.07%	44.44%	44.76%	56.11%	54.88%	+
50% ± 35%	15%	85%	44.44%	44.44%	44.44%	56.28%	55.79%	+
50% ± 36%	14%	86%	45.95%	45.95%	45.95%	56.99%	56.71%	+
50% ± 37%	13%	87%	45.33%	45.95%	45.64%	57.14%	57.62%	+
50% ± 38%	12%	88%	45.45%	46.05%	45.75%	57.22%	59.15%	+
50% ± 39%	11%	89%	44.87%	46.05%	45.45%	58.42%	61.59%	+
50% ± 40%	10%	90%	45.57%	46.75%	46.15%	58.82%	62.20%	+
50% ± 41%	9%	91%	46.25%	46.84%	46.54%	58.94%	63.11%	+
50% ± 42%	8%	92%	46.99%	47.56%	47.27%	58.77%	64.33%	+
50% ± 43%	7%	93%	46.51%	47.62%	47.06%	58.33%	65.85%	+
50% ± 44%	6%	94%	46.67%	48.84%	47.73%	58.18%	67.07%	+
50% ± 45%	5%	95%	44.79%	49.43%	46.99%	57.08%	68.90%	+
50% ± 46%	4%	96%	47.06%	52.17%	49.48%	57.94%	71.04%	+
50% ± 47%	3%	97%	44.86%	51.61%	48.00%	56.49%	72.87%	+
50% ± 48%	2%	98%	44.44%	51.61%	47.76%	56.25%	73.17%	+
50% ± 49%	1%	99%	43.36%	52.13%	47.34%	55.87%	75.30%	+
50% ± 50%	0%	100%	45.31%	51.33%	48.13%	55.36%	85.37%	+

Table B.18 How the performance changes when varying the PPC range (the 2^{nd} run).

Range	Lower	Upper	Precision	Recall	F1	Accuracy	Coverage	Bayes Factor
50% ± 1%	49%	51%	0.00%	0.00%	0.00%	71.43%	2.13%	+
50% ± 2%	48%	52%	40.00%	50.00%	44.44%	58.33%	3.66%	+
50% ± 3%	47%	53%	57.14%	57.14%	57.14%	62.50%	4.88%	+
50% ± 4%	46%	54%	57.14%	50.00%	53.33%	63.16%	5.79%	+
50% ± 5%	45%	55%	57.14%	44.44%	50.00%	63.64%	6.71%	+
50% ± 6%	44%	56%	55.56%	45.45%	50.00%	64.29%	8.54%	+
50% ± 7%	43%	57%	55.56%	38.46%	45.45%	62.50%	9.76%	+
50% ± 8%	42%	58%	46.15%	40.00%	42.86%	60.98%	12.50%	+
50% ± 9%	41%	59%	46.15%	37.50%	41.38%	63.04%	14.02%	+
50% ± 10%	40%	60%	50.00%	35.00%	41.18%	61.54%	15.85%	+
50% ± 11%	39%	61%	46.67%	33.33%	38.89%	59.26%	16.46%	+
50% ± 12%	38%	62%	43.75%	31.82%	36.84%	59.32%	17.99%	+
50% ± 13%	37%	63%	47.37%	36.00%	40.91%	60.61%	20.12%	+
50% ± 14%	36%	64%	47.62%	38.46%	42.55%	60.87%	21.04%	+
50% ± 15%	35%	65%	52.17%	40.00%	45.28%	61.84%	23.17%	+
50% ± 16%	34%	66%	52.00%	38.24%	44.07%	60.71%	25.61%	+
50% ± 17%	33%	67%	48.15%	38.24%	42.62%	59.77%	26.52%	+
50% ± 18%	32%	68%	48.28%	37.84%	42.42%	58.24%	27.74%	+
50% ± 19%	31%	69%	50.00%	36.59%	42.25%	58.16%	29.88%	+
50% ± 20%	30%	70%	54.55%	38.30%	45.00%	58.49%	32.32%	+
50% ± 21%	29%	71%	54.29%	39.58%	45.78%	59.82%	34.15%	+
50% ± 22%	28%	72%	51.22%	42.00%	46.15%	59.17%	36.59%	+
50% ± 23%	27%	73%	52.38%	43.14%	47.31%	61.11%	38.41%	+
50% ± 24%	26%	74%	51.06%	43.64%	47.06%	60.00%	41.16%	+
50% ± 25%	25%	75%	50.00%	43.10%	46.30%	59.15%	43.29%	+
50% ± 26%	24%	76%	50.00%	45.16%	47.46%	58.94%	46.04%	+
50% ± 27%	23%	77%	49.18%	46.15%	47.62%	58.23%	48.17%	+
50% ± 28%	22%	78%	48.39%	45.45%	46.87%	57.50%	48.78%	+
50% ± 29%	21%	79%	48.39%	44.78%	46.51%	57.41%	49.39%	+
50% ± 30%	20%	80%	48.39%	44.12%	46.15%	57.58%	50.30%	+
50% ± 31%	19%	81%	46.15%	44.12%	45.11%	57.06%	51.83%	+
50% ± 32%	18%	82%	44.93%	44.29%	44.60%	56.25%	53.66%	+
50% ± 33%	17%	83%	45.07%	44.44%	44.76%	55.87%	54.57%	+
50% ± 34%	16%	84%	45.07%	44.44%	44.76%	56.11%	54.88%	+
50% ± 35%	15%	85%	44.44%	44.44%	44.44%	56.28%	55.79%	+
50% ± 36%	14%	86%	45.95%	45.95%	45.95%	56.99%	56.71%	+
50% ± 37%	13%	87%	45.33%	45.95%	45.64%	57.14%	57.62%	+
50% ± 38%	12%	88%	45.45%	46.05%	45.75%	57.22%	59.15%	+
50% ± 39%	11%	89%	44.87%	46.05%	45.45%	58.42%	61.59%	+
50% ± 40%	10%	90%	45.57%	46.75%	46.15%	58.82%	62.20%	+
50% ± 41%	9%	91%	46.25%	46.84%	46.54%	58.94%	63.11%	+
50% ± 42%	8%	92%	46.99%	47.56%	47.27%	58.77%	64.33%	+
50% ± 43%	7%	93%	46.51%	47.62%	47.06%	58.33%	65.85%	+
50% ± 44%	6%	94%	46.67%	48.84%	47.73%	58.18%	67.07%	+
50% ± 45%	5%	95%	44.79%	49.43%	46.99%	57.08%	68.90%	+
50% ± 46%	4%	96%	47.06%	52.17%	49.48%	57.94%	71.04%	+
50% ± 47%	3%	97%	44.86%	51.61%	48.00%	56.49%	72.87%	+
50% ± 48%	2%	98%	44.44%	51.61%	47.76%	56.25%	73.17%	+
50% ± 49%	1%	99%	43.36%	52.13%	47.34%	55.87%	75.30%	+
50% ± 50%	0%	100%	45.31%	51.33%	48.13%	55.36%	85.37%	+

Table B.19 How the performance changes when varying the PPC range (the 3^{rd} run).

Range	Lower	Upper	Precision	Recall	F1	Accuracy	Coverage	Bayes Factor
50% ± 1%	49%	51%	0.00%	0.00%	0.00%	71.43%	2.13%	+
50% ± 2%	48%	52%	40.00%	50.00%	44.44%	58.33%	3.66%	+
50% ± 3%	47%	53%	57.14%	57.14%	57.14%	62.50%	4.88%	+
50% ± 4%	46%	54%	57.14%	50.00%	53.33%	63.16%	5.79%	+
50% ± 5%	45%	55%	57.14%	44.44%	50.00%	63.64%	6.71%	+
50% ± 6%	44%	56%	55.56%	45.45%	50.00%	64.29%	8.54%	+
50% ± 7%	43%	57%	55.56%	38.46%	45.45%	62.50%	9.76%	+
50% ± 8%	42%	58%	46.15%	40.00%	42.86%	60.98%	12.50%	+
50% ± 9%	41%	59%	46.15%	37.50%	41.38%	63.04%	14.02%	+
50% ± 10%	40%	60%	50.00%	35.00%	41.18%	61.54%	15.85%	+
50% ± 11%	39%	61%	46.67%	33.33%	38.89%	59.26%	16.46%	+
50% ± 12%	38%	62%	43.75%	31.82%	36.84%	59.32%	17.99%	+
50% ± 13%	37%	63%	47.37%	36.00%	40.91%	60.61%	20.12%	+
50% ± 14%	36%	64%	47.62%	38.46%	42.55%	60.87%	21.04%	+
50% ± 15%	35%	65%	52.17%	40.00%	45.28%	61.84%	23.17%	+
50% ± 16%	34%	66%	52.00%	38.24%	44.07%	60.71%	25.61%	+
50% ± 17%	33%	67%	48.15%	38.24%	42.62%	59.77%	26.52%	+
50% ± 18%	32%	68%	48.28%	37.84%	42.42%	58.24%	27.74%	+
50% ± 19%	31%	69%	50.00%	36.59%	42.25%	58.16%	29.88%	+
50% ± 20%	30%	70%	54.55%	38.30%	45.00%	58.49%	32.32%	+
50% ± 21%	29%	71%	54.29%	39.58%	45.78%	59.82%	34.15%	+
50% ± 22%	28%	72%	51.22%	42.00%	46.15%	59.17%	36.59%	+
50% ± 23%	27%	73%	52.38%	43.14%	47.31%	61.11%	38.41%	+
50% ± 24%	26%	74%	51.06%	43.64%	47.06%	60.00%	41.16%	+
50% ± 25%	25%	75%	50.00%	43.10%	46.30%	59.15%	43.29%	+
50% ± 26%	24%	76%	50.00%	45.16%	47.46%	58.94%	46.04%	+
50% ± 27%	23%	77%	49.18%	46.15%	47.62%	58.23%	48.17%	+
50% ± 28%	22%	78%	48.39%	45.45%	46.87%	57.50%	48.78%	+
50% ± 29%	21%	79%	48.39%	44.78%	46.51%	57.41%	49.39%	+
50% ± 30%	20%	80%	48.39%	44.12%	46.15%	57.58%	50.30%	+
50% ± 31%	19%	81%	46.15%	44.12%	45.11%	57.06%	51.83%	+
50% ± 32%	18%	82%	44.93%	44.29%	44.60%	56.25%	53.66%	+
50% ± 33%	17%	83%	45.07%	44.44%	44.76%	55.87%	54.57%	+
50% ± 34%	16%	84%	45.07%	44.44%	44.76%	56.11%	54.88%	+
50% ± 35%	15%	85%	44.44%	44.44%	44.44%	56.28%	55.79%	+
50% ± 36%	14%	86%	45.95%	45.95%	45.95%	56.99%	56.71%	+
50% ± 37%	13%	87%	45.33%	45.95%	45.64%	57.14%	57.62%	+
50% ± 38%	12%	88%	45.45%	46.05%	45.75%	57.22%	59.15%	+
50% ± 39%	11%	89%	44.87%	46.05%	45.45%	58.42%	61.59%	+
50% ± 40%	10%	90%	45.57%	46.75%	46.15%	58.82%	62.20%	+
50% ± 41%	9%	91%	46.25%	46.84%	46.54%	58.94%	63.11%	+
50% ± 42%	8%	92%	46.99%	47.56%	47.27%	58.77%	64.33%	+
50% ± 43%	7%	93%	46.51%	47.62%	47.06%	58.33%	65.85%	+
50% ± 44%	6%	94%	46.67%	48.84%	47.73%	58.18%	67.07%	+
50% ± 45%	5%	95%	44.79%	49.43%	46.99%	57.08%	68.90%	+
50% ± 46%	4%	96%	47.06%	52.17%	49.48%	57.94%	71.04%	+
50% ± 47%	3%	97%	44.86%	51.61%	48.00%	56.49%	72.87%	+
50% ± 48%	2%	98%	44.44%	51.61%	47.76%	56.25%	73.17%	+
50% ± 49%	1%	99%	43.36%	52.13%	47.34%	55.87%	75.30%	+
50% ± 50%	0%	100%	45.31%	51.33%	48.13%	55.36%	85.37%	+

Table B.20 How the performance changes when varying the PPC range (the 4^{th} run).

Range	Lower	Upper	Precision	Recall	F1	Accuracy	Coverage	Bayes Factor
50% ± 1%	49%	51%	0.00%	0.00%	0.00%	71.43%	2.13%	+
50% ± 2%	48%	52%	40.00%	50.00%	44.44%	58.33%	3.66%	+
50% ± 3%	47%	53%	57.14%	57.14%	57.14%	62.50%	4.88%	+
50% ± 4%	46%	54%	57.14%	50.00%	53.33%	63.16%	5.79%	+
50% ± 5%	45%	55%	57.14%	44.44%	50.00%	63.64%	6.71%	+
50% ± 6%	44%	56%	55.56%	45.45%	50.00%	64.29%	8.54%	+
50% ± 7%	43%	57%	55.56%	38.46%	45.45%	62.50%	9.76%	+
50% ± 8%	42%	58%	46.15%	40.00%	42.86%	60.98%	12.50%	+
50% ± 9%	41%	59%	46.15%	37.50%	41.38%	63.04%	14.02%	+
50% ± 10%	40%	60%	50.00%	35.00%	41.18%	61.54%	15.85%	+
50% ± 11%	39%	61%	46.67%	33.33%	38.89%	59.26%	16.46%	+
50% ± 12%	38%	62%	43.75%	31.82%	36.84%	59.32%	17.99%	+
50% ± 13%	37%	63%	47.37%	36.00%	40.91%	60.61%	20.12%	+
50% ± 14%	36%	64%	47.62%	38.46%	42.55%	60.87%	21.04%	+
50% ± 15%	35%	65%	52.17%	40.00%	45.28%	61.84%	23.17%	+
50% ± 16%	34%	66%	52.00%	38.24%	44.07%	60.71%	25.61%	+
50% ± 17%	33%	67%	48.15%	38.24%	42.62%	59.77%	26.52%	+
50% ± 18%	32%	68%	48.28%	37.84%	42.42%	58.24%	27.74%	+
50% ± 19%	31%	69%	50.00%	36.59%	42.25%	58.16%	29.88%	+
50% ± 20%	30%	70%	54.55%	38.30%	45.00%	58.49%	32.32%	+
50% ± 21%	29%	71%	54.29%	39.58%	45.78%	59.82%	34.15%	+
50% ± 22%	28%	72%	51.22%	42.00%	46.15%	59.17%	36.59%	+
50% ± 23%	27%	73%	52.38%	43.14%	47.31%	61.11%	38.41%	+
50% ± 24%	26%	74%	51.06%	43.64%	47.06%	60.00%	41.16%	+
50% ± 25%	25%	75%	50.00%	43.10%	46.30%	59.15%	43.29%	+
50% ± 26%	24%	76%	50.00%	45.16%	47.46%	58.94%	46.04%	+
50% ± 27%	23%	77%	49.18%	46.15%	47.62%	58.23%	48.17%	+
50% ± 28%	22%	78%	48.39%	45.45%	46.87%	57.50%	48.78%	+
50% ± 29%	21%	79%	48.39%	44.78%	46.51%	57.41%	49.39%	+
50% ± 30%	20%	80%	48.39%	44.12%	46.15%	57.58%	50.30%	+
50% ± 31%	19%	81%	46.15%	44.12%	45.11%	57.06%	51.83%	+
50% ± 32%	18%	82%	44.93%	44.29%	44.60%	56.25%	53.66%	+
50% ± 33%	17%	83%	45.07%	44.44%	44.76%	55.87%	54.57%	+
50% ± 34%	16%	84%	45.07%	44.44%	44.76%	56.11%	54.88%	+
50% ± 35%	15%	85%	44.44%	44.44%	44.44%	56.28%	55.79%	+
50% ± 36%	14%	86%	45.95%	45.95%	45.95%	56.99%	56.71%	+
50% ± 37%	13%	87%	45.33%	45.95%	45.64%	57.14%	57.62%	+
50% ± 38%	12%	88%	45.45%	46.05%	45.75%	57.22%	59.15%	+
50% ± 39%	11%	89%	44.87%	46.05%	45.45%	58.42%	61.59%	+
50% ± 40%	10%	90%	45.57%	46.75%	46.15%	58.82%	62.20%	+
50% ± 41%	9%	91%	46.25%	46.84%	46.54%	58.94%	63.11%	+
50% ± 42%	8%	92%	46.99%	47.56%	47.27%	58.77%	64.33%	+
50% ± 43%	7%	93%	46.51%	47.62%	47.06%	58.33%	65.85%	+
50% ± 44%	6%	94%	46.67%	48.84%	47.73%	58.18%	67.07%	+
50% ± 45%	5%	95%	44.79%	49.43%	46.99%	57.08%	68.90%	+
50% ± 46%	4%	96%	47.06%	52.17%	49.48%	57.94%	71.04%	+
50% ± 47%	3%	97%	44.86%	51.61%	48.00%	56.49%	72.87%	+
50% ± 48%	2%	98%	44.44%	51.61%	47.76%	56.25%	73.17%	+
50% ± 49%	1%	99%	43.36%	52.13%	47.34%	55.87%	75.30%	+
50% ± 50%	0%	100%	45.31%	51.33%	48.13%	55.36%	85.37%	+

Table B.21 How the performance changes when varying the PPC range (the 5^{th} run).

Range	Lower	Upper	Precision	Recall	F1	Accuracy	Coverage	Bayes Factor
50% ± 1%	49%	51%	0.00%	0.00%	0.00%	71.43%	2.13%	+
50% ± 2%	48%	52%	40.00%	50.00%	44.44%	58.33%	3.66%	+
50% ± 3%	47%	53%	57.14%	57.14%	57.14%	62.50%	4.88%	+
50% ± 4%	46%	54%	57.14%	50.00%	53.33%	63.16%	5.79%	+
50% ± 5%	45%	55%	57.14%	44.44%	50.00%	63.64%	6.71%	+
50% ± 6%	44%	56%	55.56%	45.45%	50.00%	64.29%	8.54%	+
50% ± 7%	43%	57%	55.56%	38.46%	45.45%	62.50%	9.76%	+
50% ± 8%	42%	58%	46.15%	40.00%	42.86%	60.98%	12.50%	+
50% ± 9%	41%	59%	46.15%	37.50%	41.38%	63.04%	14.02%	+
50% ± 10%	40%	60%	50.00%	35.00%	41.18%	61.54%	15.85%	+
50% ± 11%	39%	61%	46.67%	33.33%	38.89%	59.26%	16.46%	+
50% ± 12%	38%	62%	43.75%	31.82%	36.84%	59.32%	17.99%	+
50% ± 13%	37%	63%	47.37%	36.00%	40.91%	60.61%	20.12%	+
50% ± 14%	36%	64%	47.62%	38.46%	42.55%	60.87%	21.04%	+
50% ± 15%	35%	65%	52.17%	40.00%	45.28%	61.84%	23.17%	+
50% ± 16%	34%	66%	52.00%	38.24%	44.07%	60.71%	25.61%	+
50% ± 17%	33%	67%	48.15%	38.24%	42.62%	59.77%	26.52%	+
50% ± 18%	32%	68%	48.28%	37.84%	42.42%	58.24%	27.74%	+
50% ± 19%	31%	69%	50.00%	36.59%	42.25%	58.16%	29.88%	+
50% ± 20%	30%	70%	54.55%	38.30%	45.00%	58.49%	32.32%	+
50% ± 21%	29%	71%	54.29%	39.58%	45.78%	59.82%	34.15%	+
50% ± 22%	28%	72%	51.22%	42.00%	46.15%	59.17%	36.59%	+
50% ± 23%	27%	73%	52.38%	43.14%	47.31%	61.11%	38.41%	+
50% ± 24%	26%	74%	51.06%	43.64%	47.06%	60.00%	41.16%	+
50% ± 25%	25%	75%	50.00%	43.10%	46.30%	59.15%	43.29%	+
50% ± 26%	24%	76%	50.00%	45.16%	47.46%	58.94%	46.04%	+
50% ± 27%	23%	77%	49.18%	46.15%	47.62%	58.23%	48.17%	+
50% ± 28%	22%	78%	48.39%	45.45%	46.87%	57.50%	48.78%	+
50% ± 29%	21%	79%	48.39%	44.78%	46.51%	57.41%	49.39%	+
50% ± 30%	20%	80%	48.39%	44.12%	46.15%	57.58%	50.30%	+
50% ± 31%	19%	81%	46.15%	44.12%	45.11%	57.06%	51.83%	+
50% ± 32%	18%	82%	44.93%	44.29%	44.60%	56.25%	53.66%	+
50% ± 33%	17%	83%	45.07%	44.44%	44.76%	55.87%	54.57%	+
50% ± 34%	16%	84%	45.07%	44.44%	44.76%	56.11%	54.88%	+
50% ± 35%	15%	85%	44.44%	44.44%	44.44%	56.28%	55.79%	+
50% ± 36%	14%	86%	45.95%	45.95%	45.95%	56.99%	56.71%	+
50% ± 37%	13%	87%	45.33%	45.95%	45.64%	57.14%	57.62%	+
50% ± 38%	12%	88%	45.45%	46.05%	45.75%	57.22%	59.15%	+
50% ± 39%	11%	89%	44.87%	46.05%	45.45%	58.42%	61.59%	+
50% ± 40%	10%	90%	45.57%	46.75%	46.15%	58.82%	62.20%	+
50% ± 41%	9%	91%	46.25%	46.84%	46.54%	58.94%	63.11%	+
50% ± 42%	8%	92%	46.99%	47.56%	47.27%	58.77%	64.33%	+
50% ± 43%	7%	93%	46.51%	47.62%	47.06%	58.33%	65.85%	+
50% ± 44%	6%	94%	46.67%	48.84%	47.73%	58.18%	67.07%	+
50% ± 45%	5%	95%	44.79%	49.43%	46.99%	57.08%	68.90%	+
50% ± 46%	4%	96%	47.06%	52.17%	49.48%	57.94%	71.04%	+
50% ± 47%	3%	97%	44.86%	51.61%	48.00%	56.49%	72.87%	+
50% ± 48%	2%	98%	44.44%	51.61%	47.76%	56.25%	73.17%	+
50% ± 49%	1%	99%	43.36%	52.13%	47.34%	55.87%	75.30%	+
50% ± 50%	0%	100%	45.31%	51.33%	48.13%	55.36%	85.37%	+

Table B.22 How the performance changes when varying the PPC range (the 6^{th} run).

Range	Lower	Upper	Precision	Recall	F1	Accuracy	Coverage	Bayes Factor
50% ± 1%	49%	51%	0.00%	0.00%	0.00%	71.43%	2.13%	+
50% ± 2%	48%	52%	40.00%	50.00%	44.44%	58.33%	3.66%	+
50% ± 3%	47%	53%	57.14%	57.14%	57.14%	62.50%	4.88%	+
50% ± 4%	46%	54%	57.14%	50.00%	53.33%	63.16%	5.79%	+
50% ± 5%	45%	55%	57.14%	44.44%	50.00%	63.64%	6.71%	+
50% ± 6%	44%	56%	55.56%	45.45%	50.00%	64.29%	8.54%	+
50% ± 7%	43%	57%	55.56%	38.46%	45.45%	62.50%	9.76%	+
50% ± 8%	42%	58%	46.15%	40.00%	42.86%	60.98%	12.50%	+
50% ± 9%	41%	59%	46.15%	37.50%	41.38%	63.04%	14.02%	+
50% ± 10%	40%	60%	50.00%	35.00%	41.18%	61.54%	15.85%	+
50% ± 11%	39%	61%	46.67%	33.33%	38.89%	59.26%	16.46%	+
50% ± 12%	38%	62%	43.75%	31.82%	36.84%	59.32%	17.99%	+
50% ± 13%	37%	63%	47.37%	36.00%	40.91%	60.61%	20.12%	+
50% ± 14%	36%	64%	47.62%	38.46%	42.55%	60.87%	21.04%	+
50% ± 15%	35%	65%	52.17%	40.00%	45.28%	61.84%	23.17%	+
50% ± 16%	34%	66%	52.00%	38.24%	44.07%	60.71%	25.61%	+
50% ± 17%	33%	67%	48.15%	38.24%	42.62%	59.77%	26.52%	+
50% ± 18%	32%	68%	48.28%	37.84%	42.42%	58.24%	27.74%	+
50% ± 19%	31%	69%	50.00%	36.59%	42.25%	58.16%	29.88%	+
50% ± 20%	30%	70%	54.55%	38.30%	45.00%	58.49%	32.32%	+
50% ± 21%	29%	71%	54.29%	39.58%	45.78%	59.82%	34.15%	+
50% ± 22%	28%	72%	51.22%	42.00%	46.15%	59.17%	36.59%	+
50% ± 23%	27%	73%	52.38%	43.14%	47.31%	61.11%	38.41%	+
50% ± 24%	26%	74%	51.06%	43.64%	47.06%	60.00%	41.16%	+
50% ± 25%	25%	75%	50.00%	43.10%	46.30%	59.15%	43.29%	+
50% ± 26%	24%	76%	50.00%	45.16%	47.46%	58.94%	46.04%	+
50% ± 27%	23%	77%	49.18%	46.15%	47.62%	58.23%	48.17%	+
50% ± 28%	22%	78%	48.39%	45.45%	46.87%	57.50%	48.78%	+
50% ± 29%	21%	79%	48.39%	44.78%	46.51%	57.41%	49.39%	+
50% ± 30%	20%	80%	48.39%	44.12%	46.15%	57.58%	50.30%	+
50% ± 31%	19%	81%	46.15%	44.12%	45.11%	57.06%	51.83%	+
50% ± 32%	18%	82%	44.93%	44.29%	44.60%	56.25%	53.66%	+
50% ± 33%	17%	83%	45.07%	44.44%	44.76%	55.87%	54.57%	+
50% ± 34%	16%	84%	45.07%	44.44%	44.76%	56.11%	54.88%	+
50% ± 35%	15%	85%	44.44%	44.44%	44.44%	56.28%	55.79%	+
50% ± 36%	14%	86%	45.95%	45.95%	45.95%	56.99%	56.71%	+
50% ± 37%	13%	87%	45.33%	45.95%	45.64%	57.14%	57.62%	+
50% ± 38%	12%	88%	45.45%	46.05%	45.75%	57.22%	59.15%	+
50% ± 39%	11%	89%	44.87%	46.05%	45.45%	58.42%	61.59%	+
50% ± 40%	10%	90%	45.57%	46.75%	46.15%	58.82%	62.20%	+
50% ± 41%	9%	91%	46.25%	46.84%	46.54%	58.94%	63.11%	+
50% ± 42%	8%	92%	46.99%	47.56%	47.27%	58.77%	64.33%	+
50% ± 43%	7%	93%	46.51%	47.62%	47.06%	58.33%	65.85%	+
50% ± 44%	6%	94%	46.67%	48.84%	47.73%	58.18%	67.07%	+
50% ± 45%	5%	95%	44.79%	49.43%	46.99%	57.08%	68.90%	+
50% ± 46%	4%	96%	47.06%	52.17%	49.48%	57.94%	71.04%	+
50% ± 47%	3%	97%	44.86%	51.61%	48.00%	56.49%	72.87%	+
50% ± 48%	2%	98%	44.44%	51.61%	47.76%	56.25%	73.17%	+
50% ± 49%	1%	99%	43.36%	52.13%	47.34%	55.87%	75.30%	+
50% ± 50%	0%	100%	45.31%	51.33%	48.13%	55.36%	85.37%	+

Table B.23 How the performance changes when varying the PPC range (the 7^{th} run).

Range	Lower	Upper	Precision	Recall	F1	Accuracy	Coverage	Bayes Factor
50% ± 1%	49%	51%	0.00%	0.00%	0.00%	71.43%	2.13%	+
50% ± 2%	48%	52%	40.00%	50.00%	44.44%	58.33%	3.66%	+
50% ± 3%	47%	53%	57.14%	57.14%	57.14%	62.50%	4.88%	+
50% ± 4%	46%	54%	57.14%	50.00%	53.33%	63.16%	5.79%	+
50% ± 5%	45%	55%	57.14%	44.44%	50.00%	63.64%	6.71%	+
50% ± 6%	44%	56%	55.56%	45.45%	50.00%	64.29%	8.54%	+
50% ± 7%	43%	57%	55.56%	38.46%	45.45%	62.50%	9.76%	+
50% ± 8%	42%	58%	46.15%	40.00%	42.86%	60.98%	12.50%	+
50% ± 9%	41%	59%	46.15%	37.50%	41.38%	63.04%	14.02%	+
50% ± 10%	40%	60%	50.00%	35.00%	41.18%	61.54%	15.85%	+
50% ± 11%	39%	61%	46.67%	33.33%	38.89%	59.26%	16.46%	+
50% ± 12%	38%	62%	43.75%	31.82%	36.84%	59.32%	17.99%	+
50% ± 13%	37%	63%	47.37%	36.00%	40.91%	60.61%	20.12%	+
50% ± 14%	36%	64%	47.62%	38.46%	42.55%	60.87%	21.04%	+
50% ± 15%	35%	65%	52.17%	40.00%	45.28%	61.84%	23.17%	+
50% ± 16%	34%	66%	52.00%	38.24%	44.07%	60.71%	25.61%	+
50% ± 17%	33%	67%	48.15%	38.24%	42.62%	59.77%	26.52%	+
50% ± 18%	32%	68%	48.28%	37.84%	42.42%	58.24%	27.74%	+
50% ± 19%	31%	69%	50.00%	36.59%	42.25%	58.16%	29.88%	+
50% ± 20%	30%	70%	54.55%	38.30%	45.00%	58.49%	32.32%	+
50% ± 21%	29%	71%	54.29%	39.58%	45.78%	59.82%	34.15%	+
50% ± 22%	28%	72%	51.22%	42.00%	46.15%	59.17%	36.59%	+
50% ± 23%	27%	73%	52.38%	43.14%	47.31%	61.11%	38.41%	+
50% ± 24%	26%	74%	51.06%	43.64%	47.06%	60.00%	41.16%	+
50% ± 25%	25%	75%	50.00%	43.10%	46.30%	59.15%	43.29%	+
50% ± 26%	24%	76%	50.00%	45.16%	47.46%	58.94%	46.04%	+
50% ± 27%	23%	77%	49.18%	46.15%	47.62%	58.23%	48.17%	+
50% ± 28%	22%	78%	48.39%	45.45%	46.87%	57.50%	48.78%	+
50% ± 29%	21%	79%	48.39%	44.78%	46.51%	57.41%	49.39%	+
50% ± 30%	20%	80%	48.39%	44.12%	46.15%	57.58%	50.30%	+
50% ± 31%	19%	81%	46.15%	44.12%	45.11%	57.06%	51.83%	+
50% ± 32%	18%	82%	44.93%	44.29%	44.60%	56.25%	53.66%	+
50% ± 33%	17%	83%	45.07%	44.44%	44.76%	55.87%	54.57%	+
50% ± 34%	16%	84%	45.07%	44.44%	44.76%	56.11%	54.88%	+
50% ± 35%	15%	85%	44.44%	44.44%	44.44%	56.28%	55.79%	+
50% ± 36%	14%	86%	45.95%	45.95%	45.95%	56.99%	56.71%	+
50% ± 37%	13%	87%	45.33%	45.95%	45.64%	57.14%	57.62%	+
50% ± 38%	12%	88%	45.45%	46.05%	45.75%	57.22%	59.15%	+
50% ± 39%	11%	89%	44.87%	46.05%	45.45%	58.42%	61.59%	+
50% ± 40%	10%	90%	45.57%	46.75%	46.15%	58.82%	62.20%	+
50% ± 41%	9%	91%	46.25%	46.84%	46.54%	58.94%	63.11%	+
50% ± 42%	8%	92%	46.99%	47.56%	47.27%	58.77%	64.33%	+
50% ± 43%	7%	93%	46.51%	47.62%	47.06%	58.33%	65.85%	+
50% ± 44%	6%	94%	46.67%	48.84%	47.73%	58.18%	67.07%	+
50% ± 45%	5%	95%	44.79%	49.43%	46.99%	57.08%	68.90%	+
50% ± 46%	4%	96%	47.06%	52.17%	49.48%	57.94%	71.04%	+
50% ± 47%	3%	97%	44.86%	51.61%	48.00%	56.49%	72.87%	+
50% ± 48%	2%	98%	44.44%	51.61%	47.76%	56.25%	73.17%	+
50% ± 49%	1%	99%	43.36%	52.13%	47.34%	55.87%	75.30%	+
50% ± 50%	0%	100%	45.31%	51.33%	48.13%	55.36%	85.37%	+

Table B.24 How the performance changes when varying the PPC range (the 8^{th} run).

Range	Lower	Upper	Precision	Recall	F1	Accuracy	Coverage	Bayes Factor
50% ± 1%	49%	51%	0.00%	0.00%	0.00%	71.43%	2.13%	+
50% ± 2%	48%	52%	40.00%	50.00%	44.44%	58.33%	3.66%	+
50% ± 3%	47%	53%	57.14%	57.14%	57.14%	62.50%	4.88%	+
50% ± 4%	46%	54%	57.14%	50.00%	53.33%	63.16%	5.79%	+
50% ± 5%	45%	55%	57.14%	44.44%	50.00%	63.64%	6.71%	+
50% ± 6%	44%	56%	55.56%	45.45%	50.00%	64.29%	8.54%	+
50% ± 7%	43%	57%	55.56%	38.46%	45.45%	62.50%	9.76%	+
50% ± 8%	42%	58%	46.15%	40.00%	42.86%	60.98%	12.50%	+
50% ± 9%	41%	59%	46.15%	37.50%	41.38%	63.04%	14.02%	+
50% ± 10%	40%	60%	50.00%	35.00%	41.18%	61.54%	15.85%	+
50% ± 11%	39%	61%	46.67%	33.33%	38.89%	59.26%	16.46%	+
50% ± 12%	38%	62%	43.75%	31.82%	36.84%	59.32%	17.99%	+
50% ± 13%	37%	63%	47.37%	36.00%	40.91%	60.61%	20.12%	+
50% ± 14%	36%	64%	47.62%	38.46%	42.55%	60.87%	21.04%	+
50% ± 15%	35%	65%	52.17%	40.00%	45.28%	61.84%	23.17%	+
50% ± 16%	34%	66%	52.00%	38.24%	44.07%	60.71%	25.61%	+
50% ± 17%	33%	67%	48.15%	38.24%	42.62%	59.77%	26.52%	+
50% ± 18%	32%	68%	48.28%	37.84%	42.42%	58.24%	27.74%	+
50% ± 19%	31%	69%	50.00%	36.59%	42.25%	58.16%	29.88%	+
50% ± 20%	30%	70%	54.55%	38.30%	45.00%	58.49%	32.32%	+
50% ± 21%	29%	71%	54.29%	39.58%	45.78%	59.82%	34.15%	+
50% ± 22%	28%	72%	51.22%	42.00%	46.15%	59.17%	36.59%	+
50% ± 23%	27%	73%	52.38%	43.14%	47.31%	61.11%	38.41%	+
50% ± 24%	26%	74%	51.06%	43.64%	47.06%	60.00%	41.16%	+
50% ± 25%	25%	75%	50.00%	43.10%	46.30%	59.15%	43.29%	+
50% ± 26%	24%	76%	50.00%	45.16%	47.46%	58.94%	46.04%	+
50% ± 27%	23%	77%	49.18%	46.15%	47.62%	58.23%	48.17%	+
50% ± 28%	22%	78%	48.39%	45.45%	46.87%	57.50%	48.78%	+
50% ± 29%	21%	79%	48.39%	44.78%	46.51%	57.41%	49.39%	+
50% ± 30%	20%	80%	48.39%	44.12%	46.15%	57.58%	50.30%	+
50% ± 31%	19%	81%	46.15%	44.12%	45.11%	57.06%	51.83%	+
50% ± 32%	18%	82%	44.93%	44.29%	44.60%	56.25%	53.66%	+
50% ± 33%	17%	83%	45.07%	44.44%	44.76%	55.87%	54.57%	+
50% ± 34%	16%	84%	45.07%	44.44%	44.76%	56.11%	54.88%	+
50% ± 35%	15%	85%	44.44%	44.44%	44.44%	56.28%	55.79%	+
50% ± 36%	14%	86%	45.95%	45.95%	45.95%	56.99%	56.71%	+
50% ± 37%	13%	87%	45.33%	45.95%	45.64%	57.14%	57.62%	+
50% ± 38%	12%	88%	45.45%	46.05%	45.75%	57.22%	59.15%	+
50% ± 39%	11%	89%	44.87%	46.05%	45.45%	58.42%	61.59%	+
50% ± 40%	10%	90%	45.57%	46.75%	46.15%	58.82%	62.20%	+
50% ± 41%	9%	91%	46.25%	46.84%	46.54%	58.94%	63.11%	+
50% ± 42%	8%	92%	46.99%	47.56%	47.27%	58.77%	64.33%	+
50% ± 43%	7%	93%	46.51%	47.62%	47.06%	58.33%	65.85%	+
50% ± 44%	6%	94%	46.67%	48.84%	47.73%	58.18%	67.07%	+
50% ± 45%	5%	95%	44.79%	49.43%	46.99%	57.08%	68.90%	+
50% ± 46%	4%	96%	47.06%	52.17%	49.48%	57.94%	71.04%	+
50% ± 47%	3%	97%	44.86%	51.61%	48.00%	56.49%	72.87%	+
50% ± 48%	2%	98%	44.44%	51.61%	47.76%	56.25%	73.17%	+
50% ± 49%	1%	99%	43.36%	52.13%	47.34%	55.87%	75.30%	+
50% ± 50%	0%	100%	45.31%	51.33%	48.13%	55.36%	85.37%	+

Table B.25 How the performance changes when varying the PPC range (the 9^{th} run).

Range	Lower	Upper	Precision	Recall	F1	Accuracy	Coverage	Bayes Factor
50% ± 1%	49%	51%	0.00%	0.00%	0.00%	71.43%	2.13%	+
50% ± 2%	48%	52%	40.00%	50.00%	44.44%	58.33%	3.66%	+
50% ± 3%	47%	53%	57.14%	57.14%	57.14%	62.50%	4.88%	+
50% ± 4%	46%	54%	57.14%	50.00%	53.33%	63.16%	5.79%	+
50% ± 5%	45%	55%	57.14%	44.44%	50.00%	63.64%	6.71%	+
50% ± 6%	44%	56%	55.56%	45.45%	50.00%	64.29%	8.54%	+
50% ± 7%	43%	57%	55.56%	38.46%	45.45%	62.50%	9.76%	+
50% ± 8%	42%	58%	46.15%	40.00%	42.86%	60.98%	12.50%	+
50% ± 9%	41%	59%	46.15%	37.50%	41.38%	63.04%	14.02%	+
50% ± 10%	40%	60%	50.00%	35.00%	41.18%	61.54%	15.85%	+
50% ± 11%	39%	61%	46.67%	33.33%	38.89%	59.26%	16.46%	+
50% ± 12%	38%	62%	43.75%	31.82%	36.84%	59.32%	17.99%	+
50% ± 13%	37%	63%	47.37%	36.00%	40.91%	60.61%	20.12%	+
50% ± 14%	36%	64%	47.62%	38.46%	42.55%	60.87%	21.04%	+
50% ± 15%	35%	65%	52.17%	40.00%	45.28%	61.84%	23.17%	+
50% ± 16%	34%	66%	52.00%	38.24%	44.07%	60.71%	25.61%	+
50% ± 17%	33%	67%	48.15%	38.24%	42.62%	59.77%	26.52%	+
50% ± 18%	32%	68%	48.28%	37.84%	42.42%	58.24%	27.74%	+
50% ± 19%	31%	69%	50.00%	36.59%	42.25%	58.16%	29.88%	+
50% ± 20%	30%	70%	54.55%	38.30%	45.00%	58.49%	32.32%	+
50% ± 21%	29%	71%	54.29%	39.58%	45.78%	59.82%	34.15%	+
50% ± 22%	28%	72%	51.22%	42.00%	46.15%	59.17%	36.59%	+
50% ± 23%	27%	73%	52.38%	43.14%	47.31%	61.11%	38.41%	+
50% ± 24%	26%	74%	51.06%	43.64%	47.06%	60.00%	41.16%	+
50% ± 25%	25%	75%	50.00%	43.10%	46.30%	59.15%	43.29%	+
50% ± 26%	24%	76%	50.00%	45.16%	47.46%	58.94%	46.04%	+
50% ± 27%	23%	77%	49.18%	46.15%	47.62%	58.23%	48.17%	+
50% ± 28%	22%	78%	48.39%	45.45%	46.87%	57.50%	48.78%	+
50% ± 29%	21%	79%	48.39%	44.78%	46.51%	57.41%	49.39%	+
50% ± 30%	20%	80%	48.39%	44.12%	46.15%	57.58%	50.30%	+
50% ± 31%	19%	81%	46.15%	44.12%	45.11%	57.06%	51.83%	+
50% ± 32%	18%	82%	44.93%	44.29%	44.60%	56.25%	53.66%	+
50% ± 33%	17%	83%	45.07%	44.44%	44.76%	55.87%	54.57%	+
50% ± 34%	16%	84%	45.07%	44.44%	44.76%	56.11%	54.88%	+
50% ± 35%	15%	85%	44.44%	44.44%	44.44%	56.28%	55.79%	+
50% ± 36%	14%	86%	45.95%	45.95%	45.95%	56.99%	56.71%	+
50% ± 37%	13%	87%	45.33%	45.95%	45.64%	57.14%	57.62%	+
50% ± 38%	12%	88%	45.45%	46.05%	45.75%	57.22%	59.15%	+
50% ± 39%	11%	89%	44.87%	46.05%	45.45%	58.42%	61.59%	+
50% ± 40%	10%	90%	45.57%	46.75%	46.15%	58.82%	62.20%	+
50% ± 41%	9%	91%	46.25%	46.84%	46.54%	58.94%	63.11%	+
50% ± 42%	8%	92%	46.99%	47.56%	47.27%	58.77%	64.33%	+
50% ± 43%	7%	93%	46.51%	47.62%	47.06%	58.33%	65.85%	+
50% ± 44%	6%	94%	46.67%	48.84%	47.73%	58.18%	67.07%	+
50% ± 45%	5%	95%	44.79%	49.43%	46.99%	57.08%	68.90%	+
50% ± 46%	4%	96%	47.06%	52.17%	49.48%	57.94%	71.04%	+
50% ± 47%	3%	97%	44.86%	51.61%	48.00%	56.49%	72.87%	+
50% ± 48%	2%	98%	44.44%	51.61%	47.76%	56.25%	73.17%	+
50% ± 49%	1%	99%	43.36%	52.13%	47.34%	55.87%	75.30%	+
50% ± 50%	0%	100%	45.31%	51.33%	48.13%	55.36%	85.37%	+

Table B.26 How the performance changes when varying the PPC range (the 10^{th} run).

Table B.27, B.28 and B.29 show the experimental results for 2(b) Unsupervised Learning — Bayesian Framework.

Run	**Precision**	**Recall**	**F1**	**Accuracy**	**Coverage**	**Cost**
1	46.00%	51.49%	48.59%	55.49%	100.00%	–
2	46.00%	51.49%	48.59%	55.49%	100.00%	–
3	46.00%	51.49%	48.59%	55.49%	100.00%	–
4	46.00%	51.49%	48.59%	55.49%	100.00%	–
5	46.00%	51.49%	48.59%	55.49%	100.00%	–
6	46.00%	51.49%	48.59%	55.49%	100.00%	–
7	46.00%	51.49%	48.59%	55.49%	100.00%	–
8	46.00%	51.49%	48.59%	55.49%	100.00%	–
9	46.00%	51.49%	48.59%	55.49%	100.00%	–
10	46.00%	51.49%	48.59%	55.49%	100.00%	–
Average	46.00%	51.49%	48.59%	55.49%	100.00%	–
SD	(0.00%)	(0.00%)	(0.00%)	(0.00%)	(0.00%)	–

Table B.27 Results of 2(b)(i) Unsupervised Learning — Bayesian (BF−, PPC−) are shown.

Run	**Precision**	**Recall**	**F1**	**Accuracy**	**Coverage**	**Cost**
1	0.00%	0.00%	0.00%	71.43%	2.13%	$2\lambda_e + 321\lambda_r$
2	0.00%	0.00%	0.00%	71.43%	2.13%	$2\lambda_e + 321\lambda_r$
3	0.00%	0.00%	0.00%	71.43%	2.13%	$2\lambda_e + 321\lambda_r$
4	0.00%	0.00%	0.00%	71.43%	2.13%	$2\lambda_e + 321\lambda_r$
5	0.00%	0.00%	0.00%	71.43%	2.13%	$2\lambda_e + 321\lambda_r$
6	0.00%	0.00%	0.00%	71.43%	2.13%	$2\lambda_e + 321\lambda_r$
7	0.00%	0.00%	0.00%	71.43%	2.13%	$2\lambda_e + 321\lambda_r$
8	0.00%	0.00%	0.00%	71.43%	2.13%	$2\lambda_e + 321\lambda_r$
9	0.00%	0.00%	0.00%	71.43%	2.13%	$2\lambda_e + 321\lambda_r$
10	0.00%	0.00%	0.00%	71.43%	2.13%	$2\lambda_e + 321\lambda_r$
Average	0.00%	0.00%	0.00%	71.43%	2.13%	$2\lambda_e + 321\lambda_r$
SD	(0.00%)	(0.00%)	(0.00%)	(0.00%)	(0.00%)	(0.00%)

Table B.28 Results of 2(b)(ii) Unsupervised Learning — Bayesian (BF+, PPC+) are shown.

Run	Precision	Recall	F1	Accuracy	Coverage	Cost
1	45.31%	51.33%	48.13%	55.36%	85.37%	$125\lambda_e + 48\lambda_r$
2	45.31%	51.33%	48.13%	55.36%	85.37%	$125\lambda_e + 48\lambda_r$
3	45.31%	51.33%	48.13%	55.36%	85.37%	$125\lambda_e + 48\lambda_r$
4	45.31%	51.33%	48.13%	55.36%	85.37%	$125\lambda_e + 48\lambda_r$
5	45.31%	51.33%	48.13%	55.36%	85.37%	$125\lambda_e + 48\lambda_r$
6	45.31%	51.33%	48.13%	55.36%	85.37%	$125\lambda_e + 48\lambda_r$
7	45.31%	51.33%	48.13%	55.36%	85.37%	$125\lambda_e + 48\lambda_r$
8	45.31%	51.33%	48.13%	55.36%	85.37%	$125\lambda_e + 48\lambda_r$
9	45.31%	51.33%	48.13%	55.36%	85.37%	$125\lambda_e + 48\lambda_r$
10	45.31%	51.33%	48.13%	55.36%	85.37%	$125\lambda_e + 48\lambda_r$
Average	45.31%	51.33%	48.13%	55.36%	85.37%	$125\lambda_e + 48\lambda_r$
SD	(0.00%)	(0.00%)	(0.00%)	(0.00%)	(0.00%)	(0.00%)

Table B.29 Results of 2(b)(ii) Unsupervised Learning — Bayesian (BF+, PPC−) are shown.

B.2.5 2(c) Unsupervised Learning — Bayesian Framework + BERT

Run	Precision	Recall	F1	Accuracy	Coverage	Cost
1	42.62%	38.81%	40.62%	53.66%	100.00%	–
2	47.51%	64.18%	54.60%	56.40%	100.00%	–
3	49.18%	44.78%	46.88%	58.54%	100.00%	–
4	45.60%	42.54%	44.02%	55.79%	100.00%	–
5	47.65%	52.99%	50.18%	57.01%	100.00%	–
6	48.33%	43.28%	45.67%	57.93%	100.00%	–
7	46.88%	55.97%	51.02%	56.10%	100.00%	–
8	44.34%	35.07%	39.17%	55.49%	100.00%	–
9	46.67%	26.12%	33.49%	57.62%	100.00%	–
10	41.22%	45.52%	43.26%	51.22%	100.00%	–
Average	46.00%	44.93%	44.89%	55.98%	100.00%	–
SD	(2.56%)	(10.83%)	(6.22%)	(2.18%)	(0.00%)	–

Table B.30 Results of 2(c) Unsupervised Learning — Bayesian+BERT (BF+, PPC+) are shown.

Run	Precision	Recall	F1	Accuracy	Coverage	Cost
1	46.20%	54.48%	50.00%	55.49%	100.00%	–
2	46.67%	57.46%	51.51%	55.79%	100.00%	–
3	48.94%	51.49%	50.18%	58.23%	100.00%	–
4	46.88%	55.97%	51.02%	56.10%	100.00%	–
5	46.99%	58.21%	52.00%	56.10%	100.00%	–
6	47.65%	52.99%	50.18%	57.01%	100.00%	–
7	45.60%	42.54%	44.02%	55.79%	100.00%	–
8	49.18%	44.78%	46.88%	58.54%	100.00%	–
9	47.51%	64.18%	54.60%	56.40%	100.00%	–
10	42.62%	38.81%	40.62%	53.66%	100.00%	–
Average	46.82%	52.09%	49.10%	56.31%	100.00%	–
SD	(1.85%)	(7.86%)	(4.14%)	(1.39%)	(0.00%)	–

Table B.31 Results of 2(c) Unsupervised Learning — Bayesian+BERT (BF+, PPC−) are shown.

Appendix C

Hardware and Software

C.1 Code Repository

Code is hosted in Github: https://github.com/kingtaojasonng/Causal_Direction.

C.2 Hardware Requirements

Below are the hardware specifications used to run the experiments:

- **Model Name**: MacBook Pro
- **Processor Name**: Intel Core i7
- **Processor Speed**: 2.8 Ghz
- **Number of Processors**: 1
- **Total Number of Cores**: 4
- **Memory**: 16 GB

Alphabetical Index

Bayes Factor (BF), 29, 39
Bayes' Rule, 17
Bayesian Inference, 16
Bayesian Network, 55
Bidirectional Causal Relation, 3
Bidirectional Encoder Representations from Transformers (BERT), 3

Cable News Network (CNN), 12
Causal Induction, 11
Causal News Corpus (CNC), 7
Causal Relation, 1
Counterfactual, 56

Directed Acyclic Graph (DAG), 4, 10, 16

Earth Mover's Distance (EMD), 53, 55
Event Classification, 8
Expectation-Maximization (EM), 10

Graph Convolutional Network (GCN), 10

Hypothesis Testing, 23

Information Extraction (IE), 8

Knowledge Graph (KG), 2

Kullback-Leibler Divergence (KL Divergence), 55

Likelihoods, 18, 24, 53

Machine Learning (ML), 3, 11
Maximum A Posteriori (MAP), 29

Named Entity Recognition (NER), 8
Natural Language Processing (NLP), 2

Pointwise Mutual Information (PMI), 9
Posteriors, 18, 29
Prior Predictive Checks (PPC), 28
Priors, 18, 26, 35

Question Answering (QA), 1, 8

Relation Detection, 8

SemEval-2007 (Task 4), 6
SemEval-2010 (Task 8), 7
Sum of Square Error (SSE), 28, 37
Support Vector Machine (SVM), 9

Thematic Roles, 6

Word Mover's Distance (WMD), 55